Covid-19: An exploration of pathogen-related data sharing during a pandemic

A Data Management Plan created using DMPonline

Creator: Yo Yehudi

Affiliation: University of Manchester

Template: University of Manchester Generic Template

ORCID iD: 0000-0003-2705-1724

Project abstract:

The world is currently experiencing lockdown in many countries due to the COVID-19 pandemic. Many efforts to "hack" solutions, models, informative graphs, etc. have arisen in very short notice. Despite the urgency of this issue, with thousands dead and over two million people confirmed infected worldwide, there are data sources with restrictive licences preventing people from effectively working with pandemic related data. This raises the following questions to be studied systematically: In times of a pandemic or epidemic when rapid response is required, what are attitudes towards pathogen-related data sharing and data access? In particular: - Are these data licenced in a way that permits re-use and redistribution? - Are they made available in ways that are easy to download and re-use, e.g. API or bulk download, machine-readable with relevant metadata? - What response do various communities have to these restrictions?

Last modified: 14-06-2020

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Covid-19: An exploration of pathogen-related data sharing during a pandemic

Manchester Data Management Outline

- Ethics
- No - only institution involved
- Acquire new data

This project will interview researchers and research software engineers who are experiencing difficulties gaining access to restricted or difficult-to-use data sources, or who are circumventing the restrictions in some way (e.g. by using an alternative data source, encouraging data curation / submission to another source, or some other measure).
Data will be gathered via semi-structured interview discussions about their work in this area.

- University of Manchester Research Data Storage Service (Isilon)
- Other storage system (please list below)

Data gathered will be stored on a University of Manchester Macbook and backed up on research data storage.

- < 1 TB
- No
- 5 - 10 years
- Anonymised personal data
- Pseudonymised personal data
- Personal information, including signed consent forms

Given that this study researches difficulties with data sharing, interview responses may include comments about data providers that are disparaging or even reveal an intent to illegally circumvent data copyright issues. This means there is a chance some of the records will be considered sensitive.

During data gathering and analysis phase: all data will be stored on a University of Manchester dedicated laptop with an encrypted solid state disk. The laptop is password protected, has an automatic screen lock timeout after a few minutes, and will not be left unattended whilst unlocked. It is also possible that DropBox for Business (Manchester-managed) will be used for collaboration purposes with other Manchester researchers during this period.
Interviewees will be asked to agree verbally to a consent form on Select Survey. In video calls this form will be shared onscreen. Calls will be conducted via a University of Manchester Zoom account and recorded locally using Zoom's recording functions.
Recordings will not be published or stored long-term - they will be transcribed, verified by another researcher, and deleted. Only text transcriptions of the recordings will be retained in the long term.
Once data gathering and analysis phases are over, any personal data that was not earmarked to be shared by the participants will be deleted from its previous locations (e.g. deleted from the laptop and dropbox, but left on the research data storage).
Long-term data storage will be using the University of Manchester research data storage.
At publication stage, interview notes may be referred to anonymously in aggregate form - e.g. "52% of interviewees reported that..." or as an anonymous quote, e.g. "one interviewee stated 'One of our goals is to...'".

- No
- No
- No
- No

Caroline Jay

2020-05-22

Project details

This research project aims to observe attitudes towards scentific pathogen related data sharing during a pandemic, a scenario of acute need where fast but correct responses are critical. We hope this might inform future pandemic and epidemic scenarios and encourage effective data sharing techniques.

Information handling minimum controls: http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=30205
Information security classification, ownership and secure information handling SOP http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=29971
Research Data Management Policy http://documents.manchester.ac.uk/display.aspx?DocID=33802
Data Protection policy http://documents.manchester.ac.uk/display.aspx?DocID=14914
Records management policy http://documents.manchester.ac.uk/display.aspx?DocID=14916
Records retention policy http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=6514

Responsibilities and Resources

Yo Yehudi, the student performing this research, will be the primary person performing the steps of the data management.
Caroline Jay, supervisor for the student, will be the Data Custodian.

Question not answered.

Data Collection

Interviews: Interviews will take the form of video, audio, or text chat, and will be recorded. These recordings will be transcribed to text after the interviews.
Audiovisual files: will be recorded via Zoom, which produces audio and video recordings autimatically. These files will then be transcribed and tagged with metadata relating to the topics discussed, embedded within the text transcription file or alongside it in a machine-readible text format such as yaml or json. Audiovisual files will be deleted once they are transcribed and the transcription has been verified by another researcher.
Aggregated data taken from the audio and visual formats will be stored in a machine-readible format such as json or csv.
In all cases the machine-readible formats of data will facilitate analysis, whilst remaining simple enough that any text editor should be able to open and re-use the data without requiring proprietary or hard to install software.

There are two primary phases of data generation: The original data collection via semi-structured interviews, and a follow-up analysis phase.
Each time an interview is conducted, an entry will be made in the master participant list (anonymous) and the participant lookup list (which contains names and contact details of participants). Consent forms will be collected using a questionnaire in Select Survey at https://apps.mhs.manchester.ac.uk/. Select Survey results will be identified by the same unique identifier that identifies all other files associated with a given participant.
Collection process and file naming:
Files will be named with the unique identifier followed by a description of the data - e.g. 12345_video_transcription.txt or 12345_audio_codebook.tsv
Interview data will be originally recorded on a Macbook using Zoom. Each file will be named with the participant's unique identifying number code / pseudonym, and transcribed as soon as possible into text files. Example file names: 12345_video_transcription.txt or 12345_video_metadata.json
Data verification procedures:
Transcriptions and associated metadata files will be verified by a second researcher who was not present at the original interview. Once verified, all videos and audio files will be deleted.
Folder structure with sample file name entries:
ROOT:
participant_lookup_list.csv
anonymous_participant_list.csv
README.md
recordings/12345_video.mp4
recordings/12349_audio.m4a
transcriptions/12345_transcription.txt
transcriptions/12346_transcription.txt
transcriptions/12345_metadata.json

transcriptions/12349_metadata.json
aggregated_results/analysis_name.txt

Documentation and Metadata

The data will be accompanied by readme file in the root of the repository. This readme will describe the structure of the folders, and describe clearly which data were generated directly by the user and which data were as a result of further analyses. In scenarios where analyses were driven by computer code, this will be clearly indicated and there will be a link to the computer source code repository that generated the data. All computer code will be sufficently documented to allow someone unfamiliar with the project to re-run the analysis - possibly in the form of a Jupyter notebook.

Any anonymised parts of the data which are published publicly will include clear licence disclaimers making it clear what types of re-use are permissible.

Ethics and Legal Compliance

In addition to the notes in "Manchester Data Management Outline" sections 8 and 9, which deal with permission to re-use comments and interview data, participants who agree to interviews will be asked to verbally sign consent forms at the start of the interview. This will be done by reading and/or screensharing the consent forms and answering the questions as we go.

While data will be personally identifiable in the early stages of data gathering, late stages such as aggregated and analysed data results are not personally identifiable.

There is no reference to sexuality, race, gender, or political views, but there is a risk that interview respondents may express intent to ignore legal copyright requirements and/or make disparaging comments about data providers with restrictive sharing/use/re-use policies. To mitigate this risk, we will share any quotes we intent to publish with participants before we publish the quotes, to ensure they accept it being shared in anonymous form.

No data will be recorded or shared without the express consent of participants, and all data that are shared will be anonymised and (if relevant) aggregated.

Nevertheless, since this topic involves recording individuals and small amounts of personal data, the study plans will be reviewed by an ethical review board before the study commences.

Only data that are expressly open or which we have been given permission to gather will be used; any code for analysis (if any) will be licenced openly, and all data generated will be shared under creative commons licences, assuming we have permission to share it at all.

Storage and backup

Data will be stored on a University of Manchester macbook locally, and research data storage for backup. Backups will be made shortly after data are gathered.

Data will be stored on a macbook with an encrypted solid state disk. Access to the machine is password-protected and the machine is always locked when unattended, and locks on sleep/closed lid. There will be only one user with access to the Macbook.

Transfer of any personally identifiable data to the Manchester Research Data Storage facility will be performed via the Manchester VPN.

Selection and Preservation

This study aims to share data for re-use where possible, with the exception of data that must not be shared for personal privacy reasons.

All other data will be preserved for at least 5 years after publication in line with the University of Manchester's policies, but ideally indefinitely.

Data where the participant has expressly agreed to share openly will be shared and deposited in a data repository.

As above - data which has been collated to produce publishable results and data where the participant has expressly agreed to share openly will be shared and deposited in a data repository.

Data Sharing

Data which has been aggregated to produce publishable results and data where the participant has expressly agreed to share openly will be shared and deposited in a data repository such as Zenodo.

Question not answered.