Plan Overview

A Data Management Plan created using DMPonline

Title: Measuring understanding of biological data models.

Creator: Yo Yehudi

Principal Investigator: Yo Yehudi

Data Manager: Yo Yehudi

Affiliation: University of Manchester

Template: University of Manchester Generic Template

ORCID iD: 0000-0003-2705-1724

ID: 40289

Last modified: 21-02-2020

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

Measuring understanding of biological data models.

Manchester Data Management Outline

1. Is this project already funded?

- No

Will you be applying for funding from any of the following sources? If your funder isn't listed, please enter in the free text box provided.

3. Is The University of Manchester the lead institution for this project?

- Yes - leading a collaboration

Led by University of Manchester, but co-supervised by a University of Cambridge PI

4. What data will you use in this project (please select all that apply)?

- Acquire new data

Interviews and questionnaires will generate new data

5. Where will the data be stored and backed-up during the project lifetime?

Question not answered.

6. If you will be using Research Data Storage, how much storage will you require?

- < 1 TB

7. If you have a contractual agreement with a 3rd party data provider will any of the data associated with this project be sourced from, processed or stored outside of the

institutions and groups stated on your agreement?

- Not applicable

8. How long do you intend to keep your data for after the end of your project (in years)?

- > 20 years

Record retention procedures require this to be kept for 5 years after publication.
http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=6514
Where agreement has been obtained data will be open and shared indefinitely.

Questions about personal information

Personal information or personal data, the two terms are often used interchangeably, relates to identifiable living individuals. Special category personal data is more sensitive information such as medical records, ethnic background, religious beliefs, political opinions, sexual orientation and criminal convictions or offences information. If you are not using personal data then you can skip the rest of this section.

Please note that in line with data protection law (the General Data Protection Regulation and Data Protection Act 2018), personal information should only be stored in an identifiable form for as long as is necessary for the project; it should be pseudonymised (partially de-identified) and/or anonymised (completely de—identified) as soon as practically possible. You must obtain the appropriate ethical approval in order to use identifiable personal data.

9. What type of person identifying information will you be processing (please select all that apply)?

- Audio and/or video recordings
- Anonymised personal data
- Personal information
- Pseudonymised personal data

Names and contact information will not be stored associated directly alongside interview and questionnaire responses, but instead will be stored separately and associated with a unique numeric identifier also associated with an individual's responses.

10. Please provide details of how you plan to store, protect and ensure confidentiality of the participants' information as stated in the question above.

This study will include audio and/or video recordings of interview sessions, photographs and/or scans of visual results (the study will involve activities such as card-sorting and sketching), as well as questionnaires about individual personal backgrounds, primarily regarding their education and professional expertise. While this will be anonymised, it is possible that people with unique phrasing styles or with unique combinations of circumstances could be de-anonymised by a determined

individual. This will be low-risk as the interview data is regarding an individual's perception of genomic data models, and will not address sensitive personal topics.

Recordings and notes from interviews will be stored on a University of Manchester Macbook with FireVault drive encryption enabled, and backed up on University of Manchester servers. Recordings will be transcribed and stored in the same manner as the original recordings and notes. All videos and audio files will be deleted once the transcription is complete and has been verified.

Data sharing - aggregated data results: All participants will be made aware before participating in any interview sessions that their response data will be aggregated and published openly in an open access journal and data repository.

Data sharing - quotes and full transcripts / questionnaire answers: Participants will be allowed but not pressured to check a box that allows their quotes and data to be shared anonymously.

Recordings will not be published.

11. If you are storing personal information will you need to keep it beyond the end of the project?

- No

12. Sharing person identifiable information can present risks to participants' privacy, researchers and the institution. Will the participants' information (personal and/or sensitive) be shared with or accessed by anyone outside of the University of Manchester? This includes using 3rd party service providers such as cloud storage providers or survey platforms.

- No

13. If you will be sharing personal information outside of the University of Manchester will the individual or organisation you are sharing with be outside the EEA?

- No

14. Are you planning to use the personal information for future purposes such as research?

- No

15. Who will act as the data custodian or information asset owner for this study?

Caroline Jay

16. Please provide the date on which this plan was last reviewed (dd/mm/yyyy).

03/07/2019

Project details

What is the purpose of your research project?

The purpose of this project is to determine whether data modelling and data mapping skills are affected by the educational background of people who work in the intersection between biology and computer science, and if so, how it is affected.

What policies and guidelines on data management, data sharing, and data security are relevant to your research project?

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
Taking recordings of participants for research projects http://documents.manchester.ac.uk/display.aspx?DocID=38446

Responsibilities and Resources

Who will be responsible for data management?

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.

What resources will you require to deliver your plan?

Data Collection

What data will you collect or create?

Audiovisual files: will be recorded in .mov formats and audio in .aifc. 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.

Images sketched by users will be stored as png or jpg, with digitised versions stored in machine readible formats such as yaml or json.

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.

How will the data be collected or created?

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). Their background details 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, and exported into .csv format when the analysis phase begins.

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 QuickTime or PhotoBooth software. 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

Users may also wish to sketch or draw their responses to some of the questions. These pages will be digitised via scanner or the Macbook camera if quality permits. Data from the sketches is likely to be in list or graph format and will be converted into textual representations where possible, stored in text files. Example file names: 12345_sketch_1.jpg or 12345_sketch_1_transcription.json

Once the data gathering and transcription / conversion into text is complete, the data will be aggregated into a single file, in a machine-readible format such as json or csv. These data will be anonymised, with no personal identifying details present.

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
video/12345_video.mov
video/12347_video.mov
audio/12346_audio.aifc
audio/12349_audio.aifc
scans/12347_sketch_1.png
scans/12347_sketch_2.png
transcriptions/12345_video_transcription.txt
transcriptions/12346_audio_transcription.txt
transcriptions/12346_sketch_1_transcription.json
processed_data/12345_video_metadata.json
processed_data/12349_audio_metadata.json
aggregated_results/analysis_name.txt

Documentation and Metadata

What documentation and metadata will accompany the data?

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

How will you manage any ethical issues?

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. In addition, the risk of inadvertent breaches or de-anonymisation are low - the data gathered are purely about an individual's educational background and perceptions of the way different biological data types are related. There is no reference to sexuality, race, gender, political views, or any other topics that are generally treated as sensitive.

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.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

Data that users have consented to share openly (anonymised and/or aggregated data only) will be shared openly and re-usable under an open licence (probably CC0 Public domain, to facilitate re-use), and people will be encouraged to credit their source if they re-use the data.

Source code used to generate any results will be open source and licenced under a permissive non-copyleft licence such as MIT.

Storage and backup

How will the data be stored and backed up?

Data will be stored on the macbook used to collect it and backed up in Research Data Storage.

How will you manage access and security?

Data will be stored on a macbook with an encrypted hard 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

Which data should be retained, shared, and/or preserved?

This study aims to share all 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 which has been aggregated to produce publishable results and anonymised data where the participant has expressly agreed to share openly will be shared and deposited in a data repository.

What is the long-term preservation plan for the dataset?

As above - 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.

Data Sharing

How will you share the data?

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.

Are any restrictions on data sharing required?

Question not answered.