Measuring understanding of biological data models.

A Data Management Plan created using DMPonline

Creator: Yo Yehudi

Affiliation: University of Manchester

Template: University of Manchester

ORCID iD: https://orcid.org/0000-0003-2705-1724

Last modified: 13-05-2019

Measuring understanding of biological data models.

Manchester Data Management Outline

- No

Question not answered.

- Yes - leading a collaboration

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

- Acquire new data

Interviews and questionnaires will generate new data

Question not answered.

- < 1 TB

Question not answered.

- 5 - 10 years

Record retention procedures require this to be kept for 5 years after publication.http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=6514

- Pseudonymised personal data
- Personal information
- Anonymised personal data
- Audio and/or video recordings
- No sensitive or 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. Once the data gathering phase is complete, the name+anonymous identifier table will be deleted, so subjects will only be known by their unique anonymous identifiers.

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 may be transcribed, in which case the transcribed data will be stored in the same manner as the original recordings and notes.

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, and will be deleted once transcribed and backed up, unless the participant expressly agrees to share their recording.

- No

- No
- No
- No

Caroline Jay

09/05/2019

Project details

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.

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

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

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.
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.

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_metadata.json
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.

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

The data will be accompanied by readme file in the root of the repository. This readme will escribe 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 parts of the data which are published publicly (anonymised and with the permission of participants) will include clear licence disclaimers making it clear what types of re-use are permissible.

Ethics and Legal Compliance

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.

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.

Data that users have consented to share openly 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

Question not answered.

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.

Selection and Preservation

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

Personally identifiable data such as interview recordings will be deleted once their transcriptions are complete and verified.

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

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.

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

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.