21COP503 Data Governance and Ethics:

Data Management Plan

Manazir Najib
B720039
MSc – Data Science
Full-time

i) Table of Contents

1. Introduction

The following report contains information and analyses of the use and collection of data from the 3 participating Universities: led by Loughborough University and accompanied by the University of Edinburgh and Cardiff University. The project given to each University is the same, which is to gather administrative data from each University, the diverse data collected will be used to create digital stories comprised of student participation in workshops and interviews. It is suggested that the students, after the visualisation of the data, will create a short film addressing the personal impact upon them relating to their experiences. The report highlights relative concepts of collection and handling of the data, access of the data and the concerns following it, and also the archiving of the data. Through the thorough discussion below, solutions are provided which complies with the UK GDPR and UK Data Protection Act, and other relevant legislations.

2. Discussion

2.1. Description of Data

2.1.1. Type of study/data

Initially, to commence the project it is fundamental that the 3 universities provide the required data for the data modelling, thus for the construction of data-driven stories. There are several data that is required, student recruitment, retention and attainment by gender, ethnicity, disability, and socio-economic background. These criteria fall under administrative data type, which is very common type of data to collect for registration to a service, in this example it's for education from universities. "Administrative data will generally be drawn from a known population and will often retrieve information from an entire population rather than a sample." [1]

2.1.2. Format and scale of data

The initial data will be mixed and can be complexed, some of it can be numerical, others can be qualitative, but ideally the raw quantitative this can be stored in a table, on Excel. However, in order to collect this data, the students must be informed for consent. They are able to opt out from participation if they please- explicit data. Likewise, visual data will be collected from workshops and interviews with the students. This will be done with the approval of each individual, likewise the digital stories created by the students will only be in public if consent has been given.

The research of the students participating will differ from each university, and the maximum will be the capacity of students from each university. Although it can be large amounts of data, if separated in two 3 excel files the students' data should be able to sustain all the required information with each participating student per row – average university participation is 1500, so 1500 rows maximum per university. However, storing digital stories will require a lot of storage space, as video files are normally larger than just tabled data. Data volume can range from lest than 1MB to 1000+ MB for imaging and videos. However total storage will fall below 1 TB. Likewise, the time period for the collection of data is subjective. As an undergraduate minimum course time is 3 years, so these data are very stable and slow moving so unlikely to change frequently if at all. However, if the research is taking longer than a year, new students will be entering the university so more data can influence the data modelling.

2.1.3. Responsibility

All the data collected and produced is done internally, and no involvement of any third parties. And the digital stories will be created by the students so still internal. The PI will be the main body responsible for the management of data. Their leadership and control of the research team should be accordance to the ESRC as well as the GDPR regulations. It is essential that the management plan

is followed by each individual researcher, and it is the responsibility of the PI to follow through on procedures. The data management plan should be reviewed and monitored in regular meetings, which should highlight the importance of privacy and pseudonymisation. Loughborough University is registered with ICO as the Data Controller so, its data protection officer (Director of Operations) is responsible for ensuring that the university processes the personal data of its employees, students, customers, suppliers and partners in accordance with the applicable data protection regulations.

2.2. Data collection / generation

2.2.1. Storage

To make sure that the data is useable and meaningful to other data scientist several measures should be taken. Ideally the data will be collected into transferable spreadsheets, interviews and video should be stored together with a transcription file. The spreadsheets besides the "information for data use, should be accompanied by information for citing and discovering" [4} of the data. Therefore, detailed descriptions and annotations should follow the data set. This helps with preparations for secondary researchers. The additional information should highlight the procedures carried out, fieldwork methods, the aim of the project, and also must "explicitly describe the meanings of variables "{3] used in tables and transcriptions.

2.2.2. File Format

Quantitative data will be stored in proprietary MS Excel or even open source like .CSV files or .TXT, these files have long-term sustainability as there are very widely used. Image, audio, and video data formats vary and is determined by camera quality, but these can be downgraded to save storage space, only if downgrading from higher quality not the other way around. The Interviews can be initially collected as digital audio recordings, in MP3 or WAV format., with a transcription done on MS Word Digital imagery can be stored in JPEG, but this format may sometimes lose detail, especially after repeated editing. Digital video data files generated will be saved as MPEG-4 format, this follows ISO specifications, and the file format is almost universal.

2.2.3. Organisation

Shallow hierarchy filing system should be used for saving the files, as deep levels of folders cause confusion and loss of data. It helps to restrict level of folder to only 3-4 layers deep and limiting the entries in each file to less than 10. A widely conventions naming system would be appropriate that clarifies the date of the data, the content of the data and what type of content is withing the file and snake casing the files is good practice, for example –20211101_DigitalStory_audio.wav, 20211101_DigitalStory_trans.pdf, 20211101_DigitalStory_image.jpg. May also be important to save files with version control, if software is updated it is important to not re-write the saved file but to update so minimising loss of data and corruption. [2]

2.2.4. Backup/ Control

Its is important to preserve the master copies of the data. This is best done if they are saved in long term digital preservations, such as open and standard formats. Data should be backed up weekly to an encrypted secure server maintained by Loughborough University, this data should only be accessible by certified personnel when being used for the project- this prevents being phished for data. It is common practice to ensure multiple copies of the data is created, generally at least 3 with one being stored off-site- in Loughborough the data is stored in Hollywell and Hazelgrove distance between them is about 1 mile. Offsite storage is on MS office data centre, located in Netherlands and Ireland. [3]

2.2.5. Tools/Software

Data can be opened in the open-sourced spreadsheets or excel files for viewing. Openly free programming tolls such as R or Python can be used to read .csv, .xml and other spreadsheet files, this can then be used to manipulate and visualise the data. Basic windows media player can be used to open MP3, MPEG-4 file for viewing of digital recordings.

2.3. Access, sharing, and re-use

2.3.1. Controller

Loughborough University as the lead will be the data processor, each individual University should be the controller of the data, but they must all mirror each other when collecting the data. It is essential that only the necessary data is collected and shared with explicit consent form each individual student. Loughborough with the full data set, will be analysing the data only, but it is the universities responsibility to keep data in an anonymised format e.g., names are not required for this exercise

2.3.2. Privacy, Ethical or Confidentiality

There are several concerns. Information about students will be collected and will be shared amongst different universities. Thousands of students so hard to keep track and keep everyone confidential. Digital videos, and audio recordings are direct identifiers, that can be exposure of privacy. If any potential risks are concentred data of individuals or whole set should be aborted. The data could be lost, or hacked into, this will be a serious invasion of privacy, can affect University reputation, which can be followed by multiple lawsuits. [4]

2.3.3. Sharing/Access

The data collected from each university for the data modelling will be shared between each other, the participants whose data will be shared must be fully informed how their information is going to be used and which parties its being shared with. With consent of students in the digital stories, those stories can then be made publicly available. As the data contains private information about individuals, it cannot be accessed unless by the participant.

2.3.4. Protection of Data

Following the GDPR guidelines to, it is important to protect the privacy of the participants of the data sets. Data minimisation is essential, as data should only be collected that will serve a purpose for the research- i.e., no names, address. Anonymising the qualitative data such as interviews and recordings, can be discussed beforehand and during editing, to what is comfortable for the participant for sharing if at all, e.g., if the person wishes not to be addressed by their name, or no videos only audio, or distort the voice even.[5] Anonymising quantitative data can be done by removing direct identifiers- names, address, age, reduce the precision of a variable [5]. It is important to re-assess any remaining disclosure risks, that can be used to identify students indirectly.

2.3.5. Intellectual Property

The copyright act will only apply to the digital stories created by the students in this project. Original work such as that will automatically make the creators of the video the copyright holders. Data that is collected by the interviews which are recorded and/or transcribed have the researcher as the copyright holder, however each interviewee is an “author of his or her recorded words”. [5] If large extracts are too be used in digital stories and data is required to be shared in public, it is acceptable as long as it is for non-commercial teaching or research purposes provided that the source, distribution, and data copyright holder have prior acknowledgement.

2.4. Archiving

2.4.1. Repository

As the lead University it is expected Loughborough to hold the data in their Research Repository to ensure that the research community have long-term access to the data. By storing data, I the repository, the project leaders can ensure the research data is migrated to new formats, platforms, and storage media in accordance with good practice. It will be securely preserved and given a citeable DOI, which grants access to shared data via persistent links. [6]

2.4.2. Software

[Please see section 2.2.5 above]

2.4.3. Preservation

The university does not keep personal data longer that it is necessary for the purpose of the research or project. It is noted that any significant research data should be archived in repositor for 10 years after the completion of the project. After the 10-year retention period, the researchers should re-evaluate the data to determine if any extension is required or not. [6] At the end, it is fully anonymised or even deleted. Records can be disposed if a student asks to do so, it will be reviewed and eventually destroyed.

3. Conclusion

The Data Management plan modelled above follows the relevant policies and legislations – GDPR, ERC and Loughborough University as the lead on this project. It is essential that the plan is followed through, to collect and use data which is required by the researchers for the project. Anonymisation is fundamental, to prevent identifications direct or preventing indirect links to students and participants for the project. Security of the data, and leadership responsibly, so data is at minimal risk of being breached or ignorantly shared and distributed. It is good practice for PI and other team members to regular revise the plan and update all members on changes. For good team management roles must be assigned by the PI. There will obviously be a cost of this plan to take place, all the detailed plans above should fall within a realistic budget. This management and project after completion will hopefully assist future researchers and sponsors on discovering data and develop further understanding on data management and monitoring policies.

References

1. Connelly, R., Playford, C., Gayle, V. and Dibben, C., 2021. *The role of administrative data in the big data revolution in social science research*. [online] Available at: <https://www.sciencedirect.com/science/article/pii/S0049089X1630206X> [Accessed 12 November 2021].
2. Citeseerx.ist.psu.edu. 2021. [online] Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.8278&rep=rep1&type=pdf> [Accessed 12 November 2021].
3. Service, U., 2021. *Research data management — UK Data Service*. [online] UK Data Service. Available at: <https://ukdataservice.ac.uk/learning-hub/research-data-management/#document-your-data> [Accessed 12 November 2021].
4. Ico.org.uk. 2021. *The principles*. [online] Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/> [Accessed 12 November 2021].
5. 2021. [online] Available at: <https://www.gdprexplained.eu/> [Accessed 12 November 2021].
6. Lboro.ac.uk. 2021. *Archiving your data | Research Support | Loughborough University*. [online] Available at: <https://www.lboro.ac.uk/research/support/publishing/archiving-your-data/> [Accessed 12 November 2021].