Hypothetical Data Management Plan

Sebastian Lynch

November 8, 2024

1 Introduction

This document outlines a hypothetical data management plan (DMP) for the research project Criminality on Campus: Enhancing University Safety Through Understanding Criminality, funded by the ESRC and led by Loughborough University in collaboration with Cardiff University, the University of Edinburgh, and their respective student unions.

The aim of the project is to examine the nature, prevalence, and impact of criminal behaviour on UK university campuses, using a mixed-methods approach that includes administrative data, student surveys, interviews, and focus groups. The research will also explore the differential impacts of campus criminality across diverse student demographics.

This data management plan addresses key considerations such as data collection, storage, sharing, ethical compliance, anonymisation, and long-term preservation. It ensures that data handling practices support the project's goals while maintaining participant confidentiality and aligning with institutional, legal, and funder requirements.

2 Nature of the Data

1. What is the data? How and in what format will the data be collected? Is it numerical data, image data, text sequences, or modelling data?

This project will utilise both quantitative and qualitative data. Quantitative data will be acquired from student surveys as well as pre-existing administrative records, such as those belonging to participating universities and local law enforcement. Focus group discussions from participating institutions and interviews with students will also generate qualitative data. These sessions and interviews will be audio recorded and transcribed. Some workshops and interviews may be video recorded in order to assist with the creation of digital storytelling. The GDPR allows us to only process personal data that is "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (European Union, 2016, Article 5, 1(c)). The data will include special category data such as students' ethnicity, socioeconomic status, gender, and disability, which is permitted by GDPR for research purposes (European Union, 2016, Article 9, 2(j)).

2. How much data will be generated for this research?

This project will generate a substantial amount of data collected from multiple universities. It is difficult to estimate the amount of data that will be produced but we can make an educated guess. FLAC audio files can take around 10MB of storage per minute of audio (Firmansah, 2016 p.1). A significant number of interviews across all participating universities could result in 10s of GBs. MP4 files are larger and a large number of files could take up several terabytes. The SPSS file sizes will be negligible compared to the audio and video sizes. The volume will strongly depend on the level of participation and number of participating universities.

3. Over what period of time will the data be collected and how frequently is the data likely to change?

Due to the extent of this research, the data will be collected over a time span of 2 years. During this time, the data is likely to change significantly. Crime rates will vary considerably throughout this period; research shows that crime rates fluctuate seasonally (Hipp, 2004). Personal information is also likely to change over this period and as per Article 5, 1(d) (European Union, 2016), it is important that personal information remains up to date and accurate.

4. Will some or all of the data be produced by a third party? If so, where is it from?

Some of the data will be produced by third parties; the administrative data will be collected from the pre-existing databases of participating universities and local law enforcement agencies.

5. Who is responsible for managing the data? Who will ensure that the data management plan is carried out?

A data protection officer will need to be appointed due to Article 37, 1(c) (European Union, 2016) which states that a DPO must be appointed if "the core activities of the controller or the processor consist of processing on a large scale of special categories of data pursuant to Article 9 or personal data relating to criminal convictions and offences referred to in Article 10." Principal Investigators are responsible for the creation and implementation of the DMP (Loughborough University, 2016).

REFERENCES FOR SECTION 2:

- European Union (2016) General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). Official Journal of the European Union (OJ L119). Available at: https://gdpr-info.eu/
- Firmansah, L. and Setiawan, E.B., 2016, May. Data audio compression lossless FLAC format to lossy audio MP3 format with Huffman shift coding algorithm. In 2016 4th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.
- Hipp, John, Daniel Bauer, Patrick Curran, Kenneth Bollen. 2004. "Crimes of Opportunity or Crimes of Emotion? Testing Two Explanations of Seasonal Change in Crime." Social Forces. 1333-1372.
- Loughborough University (2016) Research Data Management Policy. Available here.

3 Data Documentation, Organization, and Storage

1. What documentation should be created to make the data understandable by other data scientists?

Article 30 (European Union, 2016) states: "each controller and, where applicable, the controller's representative, shall maintain a record of processing activities under its responsibility." The documentation providing information about aspects of the data is metadata. Our metadata will conform to the metadata standard called the Data Documentation Initiative (DDI), as recommended by the ESRC (2021, p.3). The metadata will provide context to the data such as its origin, collection techniques, and processing. The metadata will explain each variable and its data type for the quantitative data. For the qualitative data, such as interviews, metadata should clearly state the details of the session.

2. What file formats should be used? Do these formats conform to an open standard and/or are they proprietary?

Our quantitative, administrative data will be stored as an SPSS (.sav) file, which is proprietary. Despite being proprietary, SPSS files are widely used and the best option for quantitative data with extensive metadata. Our audio files will use Free Lossless Audio Codec (FLAC) (.flac). The FLAC format is open source. Videos will be stored as MPEG-4 (.mp4) files which conform to an open standard, although not fully open and free (Pereira, p.xxi). Text sequences like interview transcripts will be saved as plain text data (.txt) which is a stable, open and free format. All of our chosen file formats are recommended by the UK Data Service (2022).

3. What directory and file naming convention should be used?

Since a large amount of data will be collected and shared with the other controlling universities, a well organised and deep directory is essential for organising the data clearly and efficiently. A well organised directory is also essential as GDPR (European Union, 2016, Chapter 3) provides individuals eight rights to their personal data such as the right of access and the right to erasure. The directory has 4 layers. The first layer is the root, the second has 2 folders; data and documents. These folders have an additional 2 levels which further organise data. File names should be clear and meaningful to the point that they should uniquely identify a file. Where necessary, file names will include versioning and dates, in the form YYYY-MM-DD.

4. What local storage and backup procedures are recommended? Will this data require secure storage?

Article 5, 1(f) (European Union, 2016) requires "appropriate security" like secure storage due to the presence of personal data. Special category data and data on criminality are also present. Since this project is shared between 2 other universities, it is necessary to also have the data stored on a cloud service. OneDrive is the best option for these needs. For local storage/backup, encrypted data can be stored in Loughborough's secure data centres.

5. What tools or software will be required to read or view the data?

Our SPSS files require the proprietary software SPSS; however, this software is widely used, and its popularity is only increasing (Okagbue, 2021). Access should therefore not be an issue, at least for the short term. The other 3 file types are all easy to access with several free softwares capable of reading each of these files.

REFERENCES FOR SECTION 3:

- European Union (2016) General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). Official Journal of the European Union (OJ L119). Available at: https://gdpr-info.eu/
- (2021) ESRC research data policy. ESRC. Available here.
- Pereira, F.C. and Ebrahimi, T., 2002. The MPEG-4 book. Prentice Hall Professional.
- (2022) Recommended Formats. UK Data Service. Available at: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/
- Okagbue, H.I., Oguntunde, P.E., Obasi, E.C. and Akhmetshin, E.M., 2021. Trends and usage pattern of SPSS and Minitab Software in Scientific research. In Journal of Physics: Conference Series (Vol. 1734, No. 1, p. 012017). IOP Publishing.

4 Access, Sharing, and Re-use

1. Will Loughborough University be the data controller, a joint controller, or a data processor in this project?

Article 26(1) of the GDPR (European Union, 2016) states: "where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers." This means that Loughborough University is a joint data controller with Cardiff University and the University of Edinburgh as well as the ESRC. This is because these four institutions are the main decision-makers and control the purposes and means of the processing of the data.

2. What are the privacy, ethical, or confidentiality concerns related to the data?

The processing of personal data must conform to the 6 principles of the GDPR stated in Article 5 (European Union, 2016). Data revealing criminal activity such as substance abuse may lead to stigmatisation if individuals are identified, which means that the data subjects are vulnerable people (ESRC, 2023), and therefore requires further considerations. Personal information regarding students' ethnicity, socioeconomic background, gender, and disability status is special category data under GDPR Article 9 (European Union, 2016), requiring strict security measures. Participants must provide informed consent; the study's objectives and the extent of the data collected must be made clear.

3. Under what circumstances can the data be shared, when, and how?

The full, non-anonymised, data set must only be shared between the research teams from each of the leading universities using a safe and secure method such as OneDrive. The data and findings can only be shared further after all secondary and third party sources have given explicit, informed consent as described in Article 7 of the GDPR (European Union, 2016). Fully anonymised data may be shared with other researchers or for public use once the research findings have been published.

4. If necessary, what measures do you recommend to protect or anonymize the data?

GDPR Article 5, 1(f) (European Union, 2016) requires "appropriate security" to be put in place to protect against unauthorised or unlawful processing and against accidental loss. We will implement pseudonymisation so that in the case of accidental loss, personal information is still protected. Pseudonymisation involves de-identifying data using a coded reference and a key to the codes that is securely held. Whilst pseudonymisation is a useful tool, "it does not result in complete relief from GDPR obligations in the way anonymisation does" (Hintze, 2018). Furthermore, the data will be subject to encryption which adds a layer of security ensuring that data remains unreadable without the decryption key. There will also be limited access control allowing only authorised personnel to view the data. Audits will be conducted routinely, ensuring that the necessary techniques are being used consistently.

5. Who will hold the intellectual property rights for the data and other information created by the project? Will any copyrighted or licensed material be used? Does the university have permission to use/disseminate this material?

Loughborough University will jointly own the IP rights of the data generated alongside Cardiff University and the University of Edinburgh. Copyrighted material could be used in the digital storytelling component, such as stock videos or background music. If copyrighted or licensed material is used, explicit permission or licensing is required. Explicit consent for publication is needed from all sources of any copyrighted or licensed material used. Permission to publish findings and data must also be sought from third party data contributors such as administrative records from other institutions.

REFERENCES FOR SECTION 4:

- European Union (2016) General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). Official Journal of the European Union (OJ L119). Available at: https://gdpr-info.eu/
- (2023) Research with potentially vulnerable people. ESRC. Available here.
- Hintze, M. and El Emam, K., 2018. Comparing the benefits of pseudonymisation and anonymisation under the GDPR. Journal of Data Protection & Privacy, 2(2), pp.145-158.

5 Archiving

1. Should the data be archived in Loughborough University Research Repository or a subject data archive for long-term digital preservation?

The data should be archived in a secure, controlled access subject data archive. A fully anonymised version of the data could be uploaded to the Loughborough University Research Repository for public use. Once all personal information is removed, GDPR requirements no longer apply (European Union, 2016, Recital 26).

2. How should the data be prepared for digital preservation or data sharing? Will the data need to be anonymized or converted to more stable file formats?

To prepare for digital preservation, the data will be pseudonymised and encrypted. To prepare for public data sharing, the data will need to be completely anonymised; removing all personally identifiable information. We do not need to convert our file formats; mp4, FLAC and txt files are already stable formats. The SPSS files are from proprietary software, however, this data will only be archived for 10 years. SPSS files should remain accessible for this period.

3. Are software or tools needed to use the data? Will these be archived?

SPSS software is required but cannot be archived as it is proprietary. Our other file types are highly accessible via numerous different softwares. Metadata should clearly specify compatible software and version requirements for easy access.

4. Which tools or techniques can be used to keep the data safe (for example, to avoid data breaches)?

Protection against data breaches is essential; 5212 companies across the globe between November 2020 and October 2021 were impacted by data breaches (Duggineni, 2023, p.1). We have already mentioned 2 techniques, encryption and pseudonymisation. The pseudonymisation and encryption keys will be stored securely and separately from the encrypted data. This ensures that in the case of a data breach, personal information remains protected. Data breaches are usually associated with cyberattacks like malware, but physical tactics such as eavesdropping, dumpster diving, and social engineering are also significant causes of data exposure (Quinn, 2015, p.348). Human error can also lead to data breaches, so it is vitally important that all members of the research teams are well trained. It is important that all members of the research team are aware of these techniques. OneDrive will be configured such that only verified members of the research teams have access to the data. Their identity can be verified via 2-factor authentication.

5. How long should the data be retained? 3-5 years, 10 years, or forever?

Neither the GDPR nor ESRC provide specific timelines for the archiving of data. We will choose to retain the data for a period of 10 years. This seems sensible as it allows time for other researchers to verify findings or conduct further studies. Whilst the ESRC does not have specific guidance, other UKRI committees such as the MRC and ESPRC require minimum retention periods of 10 years and proposed retention periods beyond the minimum limit must include valid justification (UKRI, 2023, p.2).

REFERENCES FOR SECTION 5:

- European Union (2016) General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). Official Journal of the European Union (OJ L119). Available at: https://gdpr-info.eu/
- Quinn, M 2015, Ethics for the Information Age, Global Edition, Pearson Education, Limited, Harlow.

- Duggineni, S., 2023. Impact of controls on data integrity and information systems. Science and Technology, 13(2), pp.29-35.
- (2023) Retention framework for research data and records. UKRI. Available here.