Data Management Plan

(see https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf)

General Information

Administrative information

Provide information such as name of applicant, project number, funding programme, version of DMP

Project name: Test Project

Project number: ABC12345

Name of applicant: Susan Smith, +49 (0)30 2093-70072, Susan.Smith@hu-berlin.de, http://orcid.org/0000-0002-2775-6751

Project data contact: John Doe, +49 (0)30 2093-70072, John.Doe@hu-berlin.de, https://orcid.org/0000-0001-8672-0508

Funding programme: Corona Crisis and Beyond – Perspectives for Science, Scholarship and Society by VolkswagenStiftung

Version of DMP: 1.0

1 Data Description And Collection Or Re-Use of Existing Data

1a How will new data be collected or produced and/or how will existing data be re-used?

Explain which methodologies or software will be used if new data are collected or produced.

The project data will be collected via a web-based interview form. LimeSurvey hosted by Humboldt-Universität zu Berlin will be used for data collection. The data is then processed and analysed with the software R.

State any constraints on re-use of existing data if there are any.

Due to the uniqueness of the corona crisis, there is currently no comparable data that can be considered for re-analysis. A respective source research was conducted.

Explain how data provenance will be documented.

Not applicable.

Briefly state the reasons if the re-use of any existing data sources has been considered but discarded.

Not applicable.

1b What data (for example the kind, formats, and volumes), will be collected or produced?

Give details on the kind of data: for example numeric (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media.

The collected data will primarily be numeric in form of csv spreadsheets. In addition, text documents in form of txt and pdf/a documents (e.g., questionnaires) will be generated within the project. Data processing and analysis will produce R code.

Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example pdf, xls, doc, txt, or rdf).

See above.

Justify the use of certain formats. For example, decisions may be based on staff expertise within the host organisation, a preference for open formats, standards accepted by data repositories, widespread usage within the research community, or on the software or equipment that will be used.

The generated file formats are common file formats which are widely used within the scientific community. They are open, standard formats and are accepted by the data repositories that are intended for data publication (see 5a).

Give preference to open and standard formats as they facilitate sharing and long-term re-use of data (several repositories provide lists of such 'preferred formats').

See above.

Give details on the volumes (they can be expressed in storage space required (bytes), and/or in numbers of objects, files, rows, and columns).

The data volume will depend on the number of online survey participants. We expect a maximum data volume of 5 TB.

2   Documentation And Data Quality

2a   What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

Indicate which metadata will be provided to help others identify and discover the data.

The data will be described with standardized metadata standards (see below) and will get a Digital Object Identifier (DOI) for better discoverability as well as citability.

Indicate which metadata standards (for example DDI, TEI, EML, MARC, CMDI) will be used.

DataCite metadata standard will be used to ease interdisciplinary re-use. Furthermore, it is intended to describe the data with discipline-specific metadata standards, e.g. Digital Documentation Initiative (DDI), where applicable.

Use community metadata standards where these are in place.

See above.

Indicate how the data will be organised during the project, mentioning for example conventions, version control, and folder structures. Consistent, well-ordered research data will be easier to find, understand, and re-use.

A hierarchical folder structure with dedicated sub-areas for individual work packages will be created. In addition, a file naming convention that includes version control is agreed upon with the project members and documented at the beginning of the project.

Consider what other documentation is needed to enable re-use. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, and so on.

To ease re-use and improve reproducibility a readme file and codebook as well as the questionnaire and R code will be published together with the data in the data repository.

Consider how this information will be captured and where it will be recorded for example in a database with links to each item, a 'readme' text file, file headers, code books, or lab notebooks.

See above.

2b   What data quality control measures will be used?

Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, or representation with controlled vocabularies.

Data quality will be assessed e.g. with data entry validation and checks for outliers.

3   Storage And Backup During The Research Process

3a   How will data and metadata be stored and backed up during the research?

Describe where the data will be stored and backed up during research activities and how often the backup will be performed. It is recommended to store data in least at two separate locations.

The data will be stored within the institutional storage service 'HU-Box'. HU-Box offers a synchronized cloud storage with integrated version control. Data is stored at institutional servers of Humboldt-Universität zu Berlin. This ensures regular and automatic nightly backups with at least four copies at two locations in Berlin. Sensitive data will be stored separately in encrypted, password-protected folders within HU-Box. This meets the current requirements of the GDPR.

Give preference to the use of robust, managed storage with automatic backup, such as provided by IT support services of the home institution. Storing data on laptops, stand-alone hard drives, or external storage devices such as USB sticks is not recommended.

See above.

3b   How will data security and protection of sensitive data be taken care of during the research?

Explain how the data will be recovered in the event of an incident.

In case of an incident the data can be recovered from the second server or from backup.

Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships.

Only specified project members will get the password and have access to the sensitive data. Non-sensitive data will be accessible by all project members during the research. The HU-Box has a built-in access management, which will be used to control access.

Consider data protection, particularly if your data is sensitive for example containing personal data, politically sensitive information, or trade secrets. Describe the main risks and how these will be managed.

As personal data will be collected, this information will need special protection to prevent unauthorized access. Sensitive data will be anonymized as soon as possible during the research

process. Sensitive and anonymized data will be stored separately within the HU-Box. In addition, data protection will be ensured by technical measures like encryption and password protection as well as organizational measures like the minimization of persons who have access to the data.

Explain which institutional data protection policies are in place.

The HU-Box has a data protection concept and is approved by the data protection officers of Humboldt-Universität zu Berlin. In addition, the university's IT security officer monitors the service.

4 Legal And Ethical Requirements, Codes of Conduct

4a If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Ensure that when dealing with personal data protection laws (for example GDPR) are complied with:

› Gain informed consent for preservation and/or sharing of personal data.

› Consider anonymisation of personal data for preservation and/or sharing (truly anonymous data are no longer considered personal data).

› Consider pseudonymisation of personal data (the main difference with anonymisation is that pseudonymisation is reversible).

› Consider encryption which is seen as a special case of pseudonymisation (the encryption key must be stored separately from the data, for instance by a trusted third party).

› Explain whether there is a managed access procedure in place for authorised users of personal data.

An informed consent for data processing and publication of anonymized data will be gained within the online survey. The institutional data protection officers will help to formulate the consent to ensure legal compliance. Sensitive data will be anonymized as soon as possible during the research process. Sensitive and anonymized data will be stored separately within the HU-Box. The HU-Box has a built-in access management, which will be used to control access. Only specified project members will get the password and have access to the sensitive data, which is encrypted within the HU-Box.

4b How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

Explain who will be the owner of the data, meaning who will have the rights to control access:

› Explain what access conditions will apply to the data? Will the data be openly accessible, or will there be access restrictions? In the latter case, which? Consider the use of data access and re-use licenses.

› Make sure to cover these matters of rights to control access to data for multi-partner projects and multiple data owners, in the consortium agreement.

According to the data policy of Humboldt-Universität zu Berlin, the project leader Susan Smith will be responsible for data management as well as data publication. Anonymized data, documentation and R code will be made openly accessible under a Creative Commons CC BY 4.0 International license (https://creativecommons.org/licenses/by/4.0/) within a data repository (see 5a).

Indicate whether intellectual property rights (for example Database Directive, sui generis rights) are affected. If so, explain which and how will they be dealt with.

Not applicable.

Indicate whether there are any restrictions on the re-use of third-party data.

Not applicable.

4c   What ethical issues and codes of conduct are there, and how will they be taken into account?

Consider whether ethical issues can affect how data are stored and transferred, who can see or use them, and how long they are kept. Demonstrate awareness of these aspects and respective planning.

An ethics statement will be prepared and checked by the ethics committee before the project starts.

Follow the national and international codes of conducts and institutional ethical guidelines, and check if ethical review (for example by an ethics committee) is required for data collection in the research project.

See above.

5   Data Sharing And Long-Term Preservation

5a   How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Explain how the data will be discoverable and shared (for example by deposit in a trustworthy data repository, indexed in a catalogue, use of a secure data service, direct handling of data requests, or use of another mechanism).

Due to the interdisciplinary research approach and the subject-specific archiving of relevant data repositories, a diversified data publication strategy is pursued. Anonymized data, documentation and R code will be published at PsychArchives (https://www.psycharchives.org), SowiDataNet|datorium (http://data.gesis.org/sharing) or the institutional repository edoc (https://edoc.hu-berlin.de), depending on the content of the data.

Outline the plan for data preservation and give information on how long the data will be retained.

In particular, milestone versions and data underlying publications will be published and preserved for the long-term. The data will be published in trustworthy repositories mentioned above and preserved for at least ten years. Temporary files, interim results and working copies can be deleted after project end. Sensitive data will be deleted after project end.

Explain when the data will be made available. Indicate the expected timely release. Explain whether exclusive use of the data will be claimed and if so, why and for how long. Indicate whether data sharing will be postponed or restricted for example to publish, protect intellectual property, or seek patents.

The data, documentation and R code will be deposited in the respective data repository as soon as possible during the project. If necessary, an embargo period of no longer than five years after project end will be used to facilitate article publication.

Indicate who will be able to use the data. If it is necessary to restrict access to certain communities or to apply a data sharing agreement, explain how and why. Explain what action will be taken to overcome or to minimise restrictions.

The anonymized data, documentation and code will be openly accessible. An embargo might be applied.

5b   How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes.

Sensitive data need to be stored separately and destroyed after project end due to data protection reasons. In particular, milestone versions and data underlying publications will be published and preserved for the long-term. Temporary files, interim results and working copies can be deleted after project end.

Indicate how it will be decided what data to keep. Describe the data to be preserved long-term.

Non-sensitive, anonymized data and code will be published and archived for the long-term via the repositories mentioned in 5a.

Explain the foreseeable research uses (and/or users) for the data.

The data can be relevant for future research in psychology, social sciences or behavioural sciences. Especially data regarding XXX will be useful for future research in YYY.

Indicate where the data will be deposited. If no established repository is proposed, demonstrate in the data management plan that the data can be curated effectively beyond the lifetime of the grant. It is recommended to demonstrate that the repositories policies and procedures (including any metadata standards, and costs involved) have been checked.

See 5a.

5c   What methods or software tools are needed to access and use data?

Indicate whether potential users need specific tools to access and (re-)use the data. Consider the sustainability of software needed for accessing the data.

Standard text and tabular software is needed to access the data. To reproduce the analysis and use the R code, the open source software R is needed.

Indicate whether data will be shared via a repository, requests handled directly, or whether another mechanism will be used?

See 5a.

5d   How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Explain how the data might be re-used in other contexts. Persistent identifiers should be applied so that data can be reliably and efficiently located and referred to. Persistent identifiers also help to track citations and re-use.

See 5b.

Indicate whether a persistent identifier for the data will be pursued. Typically, a trustworthy, long-term repository will provide a persistent identifier.

The repositories intended for data publication will assign a Digital Object Identifier (DOI) to the data, documentation and code for better discoverability as well as citability.

6 Data Management Responsibilities And Resources

6a *Outline the roles and responsibilities for data management/stewardship activities for example data capture, metadata production, data quality, storage and backup, data archiving, and data sharing. Name responsible individual(s) where possible.*

Responsible project members for:

- data capture: John Doe

- data documentation: Individual researcher, John Doe

- metadata production: Individual researcher, Susan Smith

- data quality: Susan Smith

- ethical questions: Susan Smith

- storage and backup: Computer- and Media Service of Humboldt-Universität zu Berlin

- data archiving/sharing: Susan Smith

For collaborative projects, explain the co-ordination of data management responsibilities across partners.

Sarah Smith will coordinate data management activities between Humboldt-Universität zu Berlin and ABC University.

Indicate who is responsible for implementing the DMP, and for ensuring it is reviewed and, if necessary, revised.

Sarah Smith

Consider regular updates of the DMP.

The DMP will be regularly updated, if changes in the data management processes are necessary.

6b What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Explain how the necessary resources (for example time) to prepare the data for sharing/preservation (data curation) have been costed in. Carefully consider and justify any resources needed to deliver the data. These may include storage costs, hardware, staff time, costs of preparing data for deposit, and repository charges.

3 project month per work package and 1 student assistant position are dedicated for data management and preparation for deposit. Storage, repository and publication costs are not necessary as the intended storage location and data repositories are free of cost.

Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories. If yes, explain how much is needed and how such costs will be covered.

Not applicable.