Data Management Plan

1 Data collection and documentation
1.1 What data will you collect, observe, generate or re-use?
Questions you might want to consider:
- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?
Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated data sets. (This relates to the FAIR Data Principles F2, I3, R1 & R1.2)

We will use data from already existing datasets and collect own data.

Already existing datasets are:

A. […some data…] from the […private company…]: These are private data. We will receive the data from the years 2012-2018. The data will be accessed and stored as a csv file.
B. […some data…]: The […] data are owned by the government […] and include information on […]. The data will be accessed after signing a confidentiality contract with the government. We will access data for the years 2012-2017. The data will be retrieved and stored as csv file.
C. […some data…]: These data are owned by the government […] and include detailed information on […]. The data will be accessed after signing a confidentiality contract with the government. We will access data for the years 2012-2017. The data will be retrieved and stored as csv file.
D. […some data…]: These data are owned by the government […] and include detailed information on […]. The data will be accessed after signing a confidentiality contract with the government. We will access data for the years 2012-2017. The data will be retrieved and stored as csv file.
E. Socio-economic information (SEI) in regional level (e.g. […] income opportunities, taxation, employment, equality measures, urbanization, education): These are public datasets that are provided by the Federal Statistic Office (FSO). The data can be freely downloaded as csv files.
F. Data on […] (e.g., prices and trade regulations at the national level): These are can be accessed freely from [… indication of website…] and the FSO. The data are downloaded in a pdf version (annual report, 2012-2017) and transferred into a csv file for further analysis.
G. […some data…]: the dataset includes information on […]. The data will be accessed from the government […]. We will access data from the years 2012-2017. The data will be accessed after signing a confidentiality contract with the government. The data will be stored as a csv file.
H. Data that will be collected during the project: We will collect detailed information on […]. They will be collected during interviews with […]. The data will be stored as a csv file (text written during the interview) and mp4 files (audio recording of the interview).

We expect the entire dataset to be less than 40GB.

1.2 How will the data be collected, observed or generated? Questions you might want to consider:

- What standards, methodologies or quality assurance processes will you use?
- How will you organize your files and handle versioning?
Explain how the data will be collected, observed or generated. Describe how you plan to control and document the consistency and quality of the collected data: calibration processes, repeated measurements, data recording standards, usage of controlled vocabularies, data entry validation, data peer review, etc. Discuss how the data management will be handled during the project, mentioning for example naming conventions, version control and folder structures. (This relates to the FAIR Data Principle R1)

The data from already existing datasets including sensitive data (data A-D and G) will be accessed via a web server in an encrypted format. The password will be provided only to people working in the project. Datasets E and F will be downloaded from the respective websites.
The data will be stored at the departmental server in an encrypted version. For each dataset (A-G) an own folder will be generated with the abbreviation of the dataset (i.e. […]). Within each folder, the datasets are stored as they are provided by the deliverer (e.g. one file per year, or one single file). The file will be named accordingly (e.g. …_2012.csv).

For the dataset H will be collected during interviews. The data will be directly entered into a tablet and the interviews will potentially be recorded. An ethical approval for the survey is not requested because no health related human data will be collected. Interviewee will be informed about the aim of the study and that the information will not be given to other parties. The interviewee will be ask to sign an informed consent (one copy remains with the interviewee, one copy is for the investigators). The informed consent include the information about what data will be collected, what will be done with the data. In addition, the interviewee will be informed about their rights to not attend to the study without any consequences. The data generated during the interviews will be stored in a separate folder names "Interviews". The files are names as premisesID_interviewer_YYYYMMDD.csv (e.g. ID34_Muller_20190412.csv)

All datasets are accessed automatically so that we do not expect any wrong entries by the data collection. The only exception is the audio-recoding of the interviews. They will be transcripted twice by two different investigators to avoid errors in entries.

1.3 What documentation and metadata will you provide with the data?
Questions you might want to consider:
- What information is required for users (computer or human) to read and interpret the data in the future?
- How will you generate this documentation?
- What community standards (if any) will be used to annotate the (meta)data?
Describe all types of documentation (README files, metadata, etc.) you will provide to help secondary users to understand and reuse your data.
Metadata should at least include basic details allowing other users (computer or human) to find the data. This includes at least a name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data. Furthermore, the documentation may include details on the methodology used, information about the per-formed processing and analytical steps, variable definitions, references to vocabularies used, as well as units of measurement. Wherever possible, the documentation should follow existing community standards and guidelines. Explain how you will prepare and share this information. (This relates to the FAIR Data Principles I1, I2, I3, R1, R1.2 & R1.3)

The datasets A-G will be accessed, saved and stored only once at the time when the data owner will provide the data or the data will be downloaded from the website. Metadata will be included into the

properties tab of the csv file. In addition, a METADATA.txt file will be generated where all metadata on these datasets will be summarized.
The metadata for dataset A-G include: name for each file, the name of the person and institute who contributed to the data, the date of collection, the conditions to access the data, the confidentiality contract allocated to the data.

For the dataset H, metadata will be included into each interview csv file and into the same METADATA.txt file. The metadata will include the name for each file, the name of the collected the data, the date of collection, the conditions to collect the data (e.g. which tablet used, audio-recording taken yes-no), informed consent signed yes-no.

Throughout the project, the metadata METADATA.txt file will be extended by including information processing the original datasets (e.g. data cleaning or data analysis). Theses metadata include the file names, name of the person who processed the data, date of data process, methodology used.

It will be responsibility of each researcher to annotate data with metadata. The principal investigator will check the metadata with all involved researchers to assure data is being properly processed, documented, and stored at the time when the datasets are stored (datasets A-G) and weekly during the data collection season (dataset H). Afterwards, the PI is checking the metadata on a monthly basis.

2 Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

Questions you might want to consider:
- What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?
- Do you have the necessary permission to obtain, process, preserve and share the data? Have the people whose data you are using been informed or did they give their consent?
- What methods will you use to ensure the protection of personal or other sensitive data?

Ethical issues in research projects demand for an adaptation of re-search data management practices, e.g. how data is stored, who can access/reuse the data and how long the data is stored. Methods to manage ethical concerns may include: anonymization of data; gain approval by ethics committees; formal consent agreements. You should outline that all ethical issues in your project have been identified, including the corresponding measures in data management. (This relates to the FAIR Data Principle A1)

The data from dataset A are private and are bound by a confidentiality agreement. It will not be possible to make these data public, nor share them with third-parties. We will also be obliged to delete the data after the project has been completed.

Dataset B,C,D and G are governmental owned data, but they are also bound to a confidentiality agreement. As for dataset A, it will not be possible to share them to make these data public, nor share them with third-parties, and we will also be obliged to delete the data after the project has been completed. However, it might be possible to publish these data in an anonymized format.

Dataset E and F are public data that can be accessed freely and we are not bound to any agreement.

The data that that will be collected via interviews (dataset H) can be made available in a fully anonymized way. The interviewee will be informed in the information consent form that the data will be made available in an anonymized way.

2.2 How will data access and security be managed?

Questions you might want to consider:

- What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?
- How will you regulate data access rights/permissions to ensure the security of the data?
- How will personal or other sensitive data be handled to ensure safe data storage and -transfer?
If you work with personal or other sensitive data you should outline the security measures in order to protect the data. Please list formal standards which will be adopted in your study. An example is ISO 27001-Information security management. Furthermore, describe the main processes or facilities for storage and processing of personal or other sensitive data. (This relates to the FAIR Data Principle A1)

The datasets A-D and G will be accessed via a web server in an encrypted format. Sensitive data will be accessible only by authorized participants to the project. The list of authorized participants will be managed by the PI. Data access rules will be defined before starting the project.

All datasets will be stored on the departmental server. State-of-the-art security standards will be implemented on the servers. Our Department is protected by the University firewall, and our servers will be managed by the IT engineers of our Department. We will have full control over our data.

2.3 How will you handle copyright and Intellectual Property Rights issues?
Questions you might want to consider:
- Who will be the owner of the data?
- Which licenses will be applied to the data?
- What restrictions apply to the reuse of third-party data?
Outline the owners of the copyright and Intellectual Property Right (IPR) of all data that will be collected and generated, including the licence(s). For consortia, an IPR ownership agreement might be necessary. You should comply with relevant funder, institutional, departmental or group policies on copyright or IPR. Furthermore, clarify what permissions are required should third-party data be re-used. (This relates to the FAIR Data Principles I3 & R1.1)

The datasets A is owned by the […]. Database B, C, D and G is owned by the government […]. For all theses datasets the intellectual property rights are set out in the collaboration agreement and will remain with the owner of the data. These third-party data can only be used for the purpose of this project and we will not be allowed to re-use them again. However, we will discuss the possibility to publish the data in an anonymous way (e.g. as supplementary information within a publication).

According to regulations, the data collected by interviews (dataset H) are owned by the University of Bern. Data will be shared in an anonymous way via repository (as described in 4.1).

3 Data storage and preservation
3.1 How will your data be stored and backed-up during the research?
Questions you might want to consider:
- What are your storage capacity and where will the data be stored?
- What are the back-up procedures?
Please mention what the needs are in terms of data storage and where the data will be stored. Please consider that data storage on laptops or hard drives, for example, is risky. Storage through IT teams is safer. If external services are asked for, it is important that this does not conflict with the policy of each entity involved in the project, especially concerning the issue of sensitive data. Please specify your back-up procedure (frequency of updates, responsibilities, automatic/manual process, security measures, etc.)

The interview data will be collected by the use of a tablet. Within 24 hours, the data will be downloaded from the tablet to the departmental server. After systematic control for completeness, the data will be deleted on the tablet within 1 month after the performance of the interview. The other datasets (A-G) are stored at the departmental server. The capacity on the server attributed to our research group lies in a terabyte range. The data on the department servers is automatically backed up every night during the current week. During the past weeks, the back-up is reduced to a weekly frequency, during the past months to a monthly and for the past years to a yearly frequency.

3.2 What is your data preservation plan?
Questions you might want to consider:
- What procedures would be used to select data to be preserved?
- What file formats will be used for preservation?
Please specify which data will be retained, shared and archived after the completion of the project and the corresponding data selection procedure (e.g. long-term value, potential value for re-use, obligations to destroy some data, etc.). Please outline a long-term preservation plan for the datasets beyond the lifetime of the project. In particular, comment on the choice of file formats and the use of community standards. (This relates to the FAIR Data Principles F2 & R1.3)

According to the agreements with the […owner of the datasets…]. the data will be destroyed after the completion of the project (after publications). They will therefore not be preserved beyond the scope of the project. The interview data will be made available on a repository (see 4.1). For this purpose, we will fully anonymize the data so that it will not be possible to trace back to the premises we undertook the interview.

4. Data sharing and reuse
4.1 How and where will the data be shared?
Questions you might want to consider
- On which repository do you plan to share your data?
- How will potential users find out about your data?
Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity of data. Please also consider how the reuse of your data will be valued and acknowledged by other researchers. (This relates to the FAIR Data Principles F1, F3, F4, A1, A1.1, A1.2 & A2)

All data for potential value for re-use will be stored at BORIS Research Data. BORIS Research Data is a data repository of the University of Bern that will be accessible from 2019 onwards and free for use for researchers from the University of Bern.

4.2 Are there any necessary limitations to protect sensitive data?
Questions you might want to consider:
- Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?
Data have to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output. Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. Consider whether a non-disclosure agreement would give sufficient protection for confidential data. (This relates to the FAIR Data Principles A1 & R1.1)

As described in 2.3 and 3.2, dataset A-D and G cannot be made available. For the datasets B-D and G we will discuss how far it will be possible to share the data in an anonymous way. The dataset H will be shared at the time of the publication latest.

4.3 I will choose digital repositories that are conform to the FAIR Data Principles. [CHECK BOX]
The SNSF requires that repositories are conform to the FAIR Data Principles (Section 5 of the guidelines for re-searchers, SNSF's explanation of the FAIR Data Principles).
If there are no repositories complying with these requirements in your research field, please deposit a copy of your data on a generic platform (see examples).
If no data can be shared, this is a statement of principles.
Yes

4.4 I will choose digital repositories maintained by a non-profit organisation. [RADIO BUTTON yes/no]
If the answer is no: "Explain why you cannot share your data on a non-commercial digital repository."
The SNSF supports the use of non-commercial repositories for data sharing. Costs related to data upload are only covered for non-commercial repositories.
Yes