SNSF Data Management Plan

# Institution

ETH Zurich

# Responsibilties

Principal Investigator: Prof. Dr. Elizabeth Tilley, tilleye@ethz.ch

Data management plan contact person: Lars Schöbitz, lschoebitz@ethz.ch

# 1. Data collection and documentation

## 1.1 What data will you collect, observe, generate or re-use?

### Data types

The following types of data will be produced in this research project:

1. Personal, anonymized data (age, gender, care responsibilities, high school degree) of participants in training offered throughout the project
2. Survey data collected from sector professionals without personal details, but categorization of stakeholder types (NGO, Government, University)
3. Usage data of computing infrastructure collected from cloud-based computing services (RStudio Cloud)
4. Commit history of repositories under git version control, including personal details (name, email)
5. Tracking data of websites created as part of the project (plausible.io web analytics)
6. Currently unpublished third-party data that is re-used, processed and published in a consistent format with coherent metadata standards and data dictionaries

The origins of third-party data can be multiple, including data from NGOs, Government, and researchers at the university. The research project supports individuals in creating curated data in a consistent data structure. The curated data, including documentation and metadata will be published using the platforms GitHub and Zenodo data repository. This published and processed data will then be re-used by the research team to prepare meta-analysis about the openly published data.

Data produced in this project will be exclusively exported and stored in CSV file format. Single files will likely not exceed a file size of 50 megabytes (e.g. 25 columns, 100’000 rows).

## 1.2 How will the data be collected, observed or generated?

The subjects of data collection (persons) will be fully informed about the types of data that is collected. Data obtained during training and workshops will be collected with informed consent and participants are informed about further use of the data. They will be informed about their rights on information, data deletion and data correction.

Usage data of platforms, computing infrastructure and websites will be collected using tools for access of APIs. Unpublished data will be collected by individuals participation and through creating copies in folders using git version control.

**File organisation**

The various data types will be organised following established file naming conventions for folders and files:

* avoid capital letters
* avoid empty spaces
* use the dash “-” to connect strings in file and folder names
* use the English language
* avoid special characters
* if applicable, use ISO 8601 date format where applicable
* if applicable, place date at the beginning of the name

All data will be stored in folders (repositories) using the git version control system for tracking of changes. Each type of data will be stored in an individual repository, which follows a (project) template structure to standardize the location of raw data, processed data, and metadata. The following is an example directory tree of a repository for data:

.  
├── CITATION.cff  
├── CONDUCT.md  
├── LICENSE.md  
├── README.md  
├── data  
│ ├── README.md  
│ ├── derived-data  
│ ├── metadata  
│ └── raw-data  
├── dmp-sor4d.qmd  
├── dmp-sor4d.rmarkdown  
├── docs  
│ ├── report  
│ └── slides  
├── src  
│ └── README.md  
└── template-eth-library-dmp-snsf.docx

Each folder contains a README file in .md file format, which describes the content of the folder. The same directory structure and principles are used throughout all projects at the research group.

## 1.3 What documentation and metadata will you provide with the data?

General human-readable metadata is stored in the README file that is contained in the /metadata folder, as shown in the directory tree above. It is a template adapted from a [guide shared by Cornell University](https://data.research.cornell.edu/content/readme) and recommended for use by ETH Library under [Guidance and instructions for the ETH Zurich DMP template - Section 1: Data collection and documentation - 1.3 What documentation and metadata will you provide with the data? - Supporting resources](https://documentation.library.ethz.ch/display/DD/Data+Management+Plan+Instructions+for+ETH+Zurich+Researchers).

A machine-readable CITATION.cff following the [Citation File Format (CFF) Standard](https://citation-file-format.github.io/) will be used to document authors and contributors names and ORCID ID.

In addition to the human readable README with a description of the data, a codebook describes the variables and values, following general metadata standards (e.g. schema.org metadata standards):

| variable\_name | description |
| --- | --- |
| directory | the directory name the file is stored in (raw-data, derived-data) |
| file\_name | the name of the input data file |
| variable\_name | the name of measured variable |
| variable\_type | the type of measured variable (categorical, numeric) |
| variable\_values | if categorical, all levels. if numeric, the range |
| description | a written description of what that measured variable is |
| unit | the units the variable was measured in |

The research project aims at identifying machine-readable metadata standards that could be universally applicable to the unpublished third-party data that the projects aims at publishing openly.

Software for processing data will exclusive rely on the R statistical programming software. Cloud-computing infrastructure (i.e. RStudio Cloud) will be used to provide the users with the RStudio IDE. Versions of R and used R packages will be standardized and documented using the R package renv.

# 2. Ethics, legal and security issues

## 2.1 How will ethical issues be addressed and handled?

For sensitive/personal data collected as part of surveys and training, the necessary ethical authorizations will be obtained from the competent ethics committee. Raw data that contains confidential/sensitive personal data will be stored solely on ETH servers (Network Attached Storage, GitLab). Only the Principal Investigator and the Data Steward (internal research group employee) will be granted access to the raw data. This raw data will be collected using personal consent agreements. These agreements will clarify further use of the raw data. If at all, the raw data will be published fully anonymized and without unique identifiers that could link one data table to another. More likely, this data will only be published in an aggregated formats and as data underlying tables and figures in research products.

Please check if your project involves data relating to (in bold) one of the following ethical issues:

(yes) Human participants (This includes all kinds of human participation, incl. non-medical research, e.g. surveys, observations, tracking the location of people)

## 2.2 How will data access and security be managed?

The main concern regarding data security is the storage of personal data that may attribute data to specific persons. All personal data will be anonymized, however, unique identifiers will be created to link individual data tables. Such sensitive data will only be accessible by the Principle Investigator and the Data Steward (internal research group employee). No sensitive/personal data will be stored in cloud service external to ETH Zurich.

## 2.3 How will you handle copyright and Intellectual Property Rights issues?

All data that is non-sensitive and generated as part of the research project will be released under CC0 license. For third-party data, that is re-used and published in a processed format, the aim is to apply at least CC BY 4.0 license.

The source code for the processing of data will be released under permissive licenses (e.g. MIT License).

# 3. Data storage and preservation

## 3.1 How will your data be stored and backed-up during the research?

No data will be stored solely on the laptop of a researcher.

All data will be stored using git version control with at least one remote repository:

1. GitLab for sensitive data
2. GitHub for non-sensitive data

## 3.2 What is your data preservation plan?

All generated data that is published openly using the long-term archiving service Zenodo will be stored for an unlimited time period. Sensitive data that is stored on institutional infrastructure (e.g. Network Attached Storage) will be archived and stored for 10 years.

All re-used third-party data with high quality metadata will be stored using long-term archiving service Zenodo an unlimited time period.

At the end of the project, the Data Steward will remain responsible for metadata improvement, integrity check and measured to ensure accessibility.

# 4. Data sharing and reuse

## 4.1 How and where will the data be shared?

The GitHub Zenodo integration will be used for sharing of data. This includes all generated data and data underlying reports and publications.

## 4.2 Are there any necessary limitations to protect sensitive data?

All non-sensitive data will be published prior to the publication of research products. Data will be treated as a product in itself. Sensitive data will not be published at any time, but only as aggregated data underlying research products and the time of publication. If sensitive data can be sufficiently anonymized, we will publish it alongside the derived aggregated data. To ensure computational reproducibility, we will also publish the scripts that contain programming code which reproduce findings from the given raw data.

## 4.3 All digital repositories I will choose are conform to the FAIR Data Principles

(yes)

## 4.4 I will choose digital repositories maintained by a non-profit organisation

(yes)