EPFL Data Management Plan (DMP) Template

Goal of a Data Management Plan (DMP)

The main objective of a DMP is to anticipate the project needs and requirements in terms of:

- resources (e.g., servers, hard drives, data curation and preservation, software tools, etc.)
- good practices (e.g., standardized documentation, metadata sharing, naming convention, data security, open formats, regular backup, etc.)
- data protection and ethics authorizations (e.g., submission to the cantonal ethics commission, or to the HREC)

A DMP is a live document supposed to be updated to reflect the actual data management. At the end of a research project, the associated DMP becomes a precious document that enables research reproducibility, allowing other researchers to understand your processes.

Scope of this document

This DMP template applies to any EPFL research project that produces, collects or processes research data that does not already have a DMP from a funder or publisher. Develop a single DMP for each project to cover its overall approach. When specific issues arise for individual datasets (e.g., openness), clearly identify them. However, depending on the project and research field, some of those issues may not need to be addressed in the DMP.

How to use of this document – READ CAREFULLY

Throughout the DMP, the required information is indicated as Mandatory [M], Mandatory if applicable [MA], or Recommended [R]. In the fields for the answers, if present, replace the text in italic with your own answer or description.

- For DMP review or relevant information on the DMP: contact researchdata@epfl.ch
- For data privacy issues and information on EPFL privacy policy: contact dpo@epfl.ch
- For ethics authorizations or funding information: contact research@epfl.ch
- For industrial collaboration and patent issues: contact tto@epfl.ch

If sharing this DMP file with the previous services, please save/rename its shared copy as: *SURNAME_Firstname_DMP.odt*. All personal information contained in the DMP will be treated as confidential, as per EPFL Privacy Policy.

Part 1. General Information

[M] Personal Information

P1a. Fellow/PI's full name(s) and email(s)	Franziska Zollner, franziska.zollner@epfl.ch Michaël Aklin, michael.aklin@epfl.ch
P1b. Fellow/PI's ORCID(s)	https://orcid.org/0009-0005-2766-1890 https://orcid.org/0000-0003-3912-082X
P1c. DMP contact person's full name and email	Franziska Zollner, franziska.zollner@epfl.ch

[M] Project and Fellowship

P1d. Project title	Impact of the Tanzanian Standard Gauge Rail on Economic Development
P1e. Project description (short, max 2-3 lines)	Using a dataset of mobile phone location data, this project evaluates the impact of the Tanzanian Standard Gauge Railway on regional mobility flows. Coupled with publicly available remote sensing data on built-up, this project will analyze how public transport shapes urban development.
P1f. Name of fellowship/funding programme	Funding of Chair
P1g. Institution(s) involved	EPFL
P1h. Laboratory(s) involved	PASU
P1i. Starting date of the project (estimated if unknown)	01.03.2026
P1j. How much of the funding will be allocated for data management activities? (Estimated if unknown)	200 CHF (40CHF/Tb/year of storage)

[M] Global context of Research Data

P1k. Could any of the data/code that you create, produce, collect or process contain personal data[1]?	x Yes *(If "Yes", then fill in Annex 2: Data Protection Analysis)* ☐ No
P1l. Could any of the data that you collect, or process contain sensitive data[2] or data that you are legally obliged to protect?	☐ Yes *(If "Yes", then fill in Annex 2: Data Protection Analysis)* x No

[1] Personal data: information that relates to an identified or identifiable individual. Not only data that can directly identify a person (e.g., name) is considered personal data, but also data that can make a person identifiable through the combination of data (e.g., combining age, e-mail address and information related to usage of social networks may allow identifying a person). You should consider that information together with all the means reasonably likely to be used by either you or any other person to identify that individual. See also **Art. 3**.c of the FADP Swiss law.

[2] Sensitive data: information that includes but is not limited to religious, ideological, political or trade union-related views or activities; health, genetic, biometric, or concerning the intimate sphere or the racial origin; ethnic data or social security measures; administrative or criminal proceedings and sanctions. See also **Art. 3**.c of the FADP Swiss law.

P1m. Could any of the data that you collect, or process contain confidential data or intellectual property of a third party that you are legally obliged to protect?	☐ Yes x No
P1n. Will you publish any of your data/code (Open Research Data)[3]?	☐ Yes *(if "Yes", then fill in Annex 1: Open Research Data/Code (ORD))* x No
P1o. P1o. Are ethics authorizations necessary[4] from the competent ethics committee (HREC, AREC)?	x Yes (If "Yes", please check: ☐ Obtained x Not obtained) ☐ No

[3] Either because of an obligation (e.g., funder or publisher requirement), because it is the norm in your community, or because you are committed to doing so yourself.

[4] Please, check the compliance in the EPFL Research ethics webpage: www.epfl.ch/research/ethic-statement/compliance

Part 2. Data & Code Collection[5]

Specify the type of data and code that is acquired, produced, processed or coded during the project, and describe them. Add more rows if necessary. *Text in italics* should be deleted.

P2a. [M] Dataset identifier #	**P2b. [M]** Collection[5] method	**P2c. [M]** Description of dataset, type of data/code	**P2d. [M]** Origin / source / equipment	**P2e. [M]** Formats	**P2f. [R]** Description of metadata	**P2g. [R]** Naming convention, format	**P2h. [R]** Person in charge
1	Observational data, acquired	Mobile Phone GPS data	Acquired from third party (Quadrant)	To be determined, likely .csv			Franziska Zollner

P2i [M]. Does any of the dataset contains personal data?	x Yes *(If yes, specify which dataset(s) and complete Annex 2: Data Protection Analysis)* ☐ No

[5] Acquired, produced, processed, or coded.

Specify the dataset(s) containing personal data *Note: ensure Annex 2: Data Protection Analysis is filled out accordingly.*	GPS mobile phone data acquired from Quadrant

Understandability of Acquired, Produced, Processed or Coded Data

Describe how you make your datasets understandable for yourself and for others. *Text in italics* should be deleted.

P2j.[R] For whom are you collecting[5] this data/code?	The data is acquired for research purposes for the PhD thesis of Franziska Zollner
P2k. [MA] What documentation do you provide with your data/code for your audience to make data/code understandable for future reuse?	The data will be confidential and will not be shared with anyone. Hence, no specific documentation will be available
P2l. [R] How will you ensure the quality of your data/code?	They data will be checked to see whether it is coherent (e.g., checking for “teleporting behavior” of mobile devices) and complete. The code used to do these quality checks will be discussed with the student’s supervisor, Michaël Aklin

Part 3. Data & Code Storage or Management Infrastructure

For column *P3g*, refer to the dataset identifiers # in *P2a* when discussing the same data or code. Add new rows as necessary. *Text in italics* should be deleted.

P3a. [M] Storage or management system	**P3b. [M]** Role of system	**P3c. [M]** Backup strategy	**P3d. [M]** System provider(s)	**P3e. [M]** Access rights management	**P3f. [M]** Estimated size of dataset	**P3g. [MA]** Associated datasets (# from P2a)
The data collected in this project will be on servers physically located at the EPFL campus on a NAS (Network Attached Storage). This is a centrally offered storage system that provides redundancy, replication, ransomware protection, controls integrity and uses the identity management system of EPFL.	Only necessary data required as input will be transferred to the compute environment to run the analysis. The compute environment is on-premises and is operated by the EPFL Research Computing Platform. Access rights are also managed based on the EPFL identity management system. Once the analysis has been run, necessary output will be recopied on the NAS.	The NAS system has regular snapshots.	The NAS is physically in the EPFL datacenters, and the service is provided internally by the Research Computing Platform	Access rights are monitored through the EPFL identity management system. Only project members have access to the NAS system through the local network with their login credentials	About 1 TB	1
RCP storage	This storage will be used during analysis of the data.	Backup strategy of RCP clusters	The RCP is physically on EPFL campus, and the service is provided internally by the Research Computing Platform	Access rights are monitored through the EPFL identity management system. Only project members have access to the RCP system through the local network with their login credentials	About 1 TB	1

Part 4. Data & Code Long Term Preservation[6]

It is expected that the preservation for all datasets of *Part 2* is discussed here. For column *P4a*, use all and the same identifiers # as in *P2a*.

P4a. [MA] # (same as P2a.)	**P4b. [M]** What will be preserved?	**P4c. [M]** Preservation purpose	**P4d. [M]** Retention Period	**P4e. [M]** Deletion process of the preserved data	**P4f. [MA]** Preservation infrastructure	**P4g. [R]** When/How often will preservation be done?	**P4h. [M]** Person in charge
1	The whole dataset will be preserved	The data will be kept until any related paper/thesis is published and for 5 years after the PhD thesis submission in case anyone challenges the published results	5 years after the PhD thesis submission of Franziska Zollner	There will be an automated, secure deletion process	ACOUA		Franziska Zollner, Michaël Aklin

[R] How is preservation achieved

[6] Digital preservation can be defined as a "series of managed activities necessary to ensure continued access to digital materials for as long as necessary" (Digital Preservation Handbook).

P4i. What data curation process(es) need(s) to be applied to your data/code (whether it will be published or not) to make them reusable in the long run?	A README will be created that describes the data and code used for this research project. The datasets will be cleaned from void or corrupted files or folders. The folders will follow a harmonized naming convention.

Annex 1. Open Research Data & Code (ORD)

Use this part when datasets are disseminated[7] in a publicly findable platform or repository[8]. The EPFL recommends following the principle that “Data should be as open as possible, as restricted as necessary” (Karel Luyben, President EOSC Association).

[7] Some types of data (e.g., non-anonymized, relevant for patent, double use technology, …) cannot be published: it is recommended to publish at least the description of those data.

[8] For definitions and suggestions of data dissemination platforms, check the online comparative table at go.epfl.ch/datarepo.

[M] Intended Purpose of Opening Up the Datasets

A1a. What is the intended purpose of opening your data/code?	

List of All Datasets (Data and Code) That Will Be Opened, Where and How

For column *A1b*, use all and the same identifiers # as in *P2a*.

A1b. [MA] Associated datasets (# from P2a)	**A1c. [M]** Title	**A1d. [M]** Author(s)	**A1e. [M]** Type/Formats	**A1f. [M]** Repository/Platform	**A1g. [M]** License	**A1h. [MA]** Anonymization process[9], if necessary

[R] Engaging the Research Community in Reusing your Dataset(s)

[9] Pseudonymized: the identifying information (surname, date of birth, ...) is replaced by a code (e.g., hashing); the key to recover the information, while kept secure, can still be used to identify the person. Anonymized: the personal identity is sufficiently protected by reducing the information, preventing anyone from re-identifying the person, unless they go to great lengths to do so.

A1i. What actions (if any) do you plan to promote or facilitate the reuse of published open data?	

Annex 2. Data Protection Analysis

[M] Data and Datasets Collected and/or Produced

A2a. How many participants will be included??	For a week of mobile phone location data, there are around 200 000 users. I intend to acquire a year of data which will likely result in a higher number of participants.	
A2b. Participants' geographical location: do you consider only Swiss residents? UE residents? Others?	I don't have information about the residence of the participants, but I only use mobile phone data from users observed in Tanzania	
A2c. Consent: have you planned to get or already obtained the subjects' consent?	x Yes ☐ No	Mobile apps obtain opt-in consent directly from users, usually during the first-use launch.

[M] Assessment Of 'Probably High-Risk' Data Processing (9 Criteria)

Under the current FADP (in force since September 1, 2023), the data controller (i.e., the person who determines the purpose and means of data processing) must conduct a Data Protection Impact Assessment (DPIA) if the processing poses high risks to data subjects. Additionally, if the GDPR applies to your project and two or more of the following criteria are met, a DPIA is required.

A2d. (1) The data processing involves: Evaluation or scoring, including profiling and predicting[10]	☐ Yes x No	

[10] Especially important if it involves "aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behavior, location or movements"

A2e. (2) The data processing involves: Automated decision making with legal or similar significant effect[11]	☐ Yes x No	
A2f. (3) The data processing involves: Systematic monitoring to observe, monitor or control data subjects[12]	☐ Yes x No	

[11] Processing that aims at taking decisions on data subjects producing “legal effects concerning the natural person” or which “similarly significantly affects the natural person”

[12] Including data collected through networks or “a systematic monitoring of a publicly accessible area”

A2g. (4) The data processing involves: Personal or sensitive data	x Yes ☐ No	The data includes variables such as device ID, IP address, geolocation and timestamps etc. which qualify as personal data
A2h. (5) The data processing involves: Data processed on a large scale	x Yes ☐ No	Yes, mobile phone data is gathered and aggregated from multiple thousand users
A2i. (6) The data processing involves: Matching or combining datasets	☐ Yes x No	
A2j. (7) The data processing involves: Data concerning vulnerable subjects[13] (e.g., children, adults lacking capacity, pregnant women, …)	x Yes ☐ No	As there is no particular target group, vulnerable subjects might be involved if they own a device that is included in the dataset. There is no way for me to verify the 'vulnerability' status of subjects. As mentioned in point A2m., I will not be actively looking for fingerprints of vulnerable population's segments.

[13] Vulnerability situations may involve: physical or cognitive impairment; subordination to an authority figure or formal authority; social discrimination; stigmatization; unfavorable socio-economic conditions; increased physical or psychological sensitivity to interventions planned in research. The Swiss law defines certain people as particularly vulnerable, thus requiring special protection.

A2k. (8) The data processing involves: Innovative use or applying new technological or organizational solutions	☐ Yes x No	
A2l. (9)The processing in itself "prevents data subjects from exercising a right or using a service or a contract"	☐ Yes x No	

[M] Security Measures

A2m. Is the data processed in a manner that ensures appropriate security[14] of the personal data using appropriate technical or organizational measures?	x Yes ☐ No	Yes, the data will not be used to identify or track any individual, it will only be used to calculate aggregate mobility statistics
A2n. Which technical and organizational measures do you plan to ensure IT security?	The data will only be shared with a very limited number of people. Access rights are monitored through the EPFL identity management system. Only project members (i.e., Franziska Zollner, PhD student, and Michaël Aklin, supervisor) have access to the NAS system through the EPFL local network with their login credentials	

[14] This might include protection against unauthorized or unlawful processing and against accidental loss, destruction, or damage.

A2o. How do you ensure that technical and organizational measures are appropriate and correctly implemented regarding the risks?	This will be handled by EPFL IT that manage the RCP facility.	
A2p. Do you plan to share personal research data with another EPFL entity?	☐ Yes x No	
A2q. Do you plan to share personal research data with anyone outside of EPFL?	☐ Yes x No	
A2r. Do you plan to work with a data processor[15]?	☐ Yes x No	
A2s. Do you have a process in place for granting and revoking appropriate user access (incl. privileged access rights and roles)?	x Yes ☐ No	Access rights are monitored through the EPFL identity management system. Only project members have access to the NAS system through the local network with their login credentials. Access can be revoked by contacting EPFL services.

[15] E.g., private company or other entity that carries out the data processing on the behalf of your research group or laboratory.

A2t. Do you have standards for isolating sensitive data and procedures or technologies in place to protect it from unauthorized access and tampering?	☐ Yes x No	No. The access to the data will be limited to a selected group of people (as mentioned before).
A2u. Do you plan to implement Privacy-by-Design or by-Default[16]?	x Yes ☐ No	Privacy-by-design. The measures are limiting the access to the data to a few selected people. Furthermore, the raw data set will be safely stored and the data set used for daily analysis will undergo several anonymization techniques. GPS coordinates will be truncated, device ID will be hashed, variables that are not necessary for my research question will be deleted.
A2v. How do you plan to ensure the traceability of personal data?[17]	I will create a document (logs) that reports who accessed/created/modified/exported/transferred/deleted which personal data, when, and how. I will keep these logs for at least two years, separate from the data themselves.	

[16] Privacy by Designs: if personal or sensitive data are involved, privacy measures are taken since the initial research stages and throughout its complete development. Privacy by default: if the research includes choices for individuals on how much personal data they share, the default settings should be the most privacy friendly ones. See also https://gdpr.eu/article-25-data-protection-by-design.

[17] Within the scope of the new FADP, you must keep the event logs related to personal data processing for a period of 2 years.