CLARIN-CH

SWISS RESEARCH DATA SUPPORT NETWORK

Data Management Plans for Language Data

Workshop

PART I – WHY DMPS MATTER FROM THE START
4 MAY, 1:00–2:30 PM (ONLINE)

PART II – HOW TO PREPARE A GOOD DMP
29 JUNE, 1:00–2:30 PM (ONLINE)

KEYNOTE & THEORETICAL INPUT | HANDS-ON ACTIVITIES

2. Legal and ethical aspects

Suzanna Farace Marazza

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 research 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)
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)
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 reused. (This relates to the *FAIR Data Principles* I3 & R1.1)

Schweizerischer Nationalfonds
Fonds national suisse
Fondo nazionale svizzero
Swiss National Science Foundation

https://www.snf.ch/media/en/4i9AE5YElf7tqhGz/DMP_content_mySNF-form_en.pdf

2. Legal and ethical aspects

Suzanna Farace Marazza

Common Privacy and Ethical Issues

- Confusion between **research data** and **personal data**; and between personal data and **sensitive data** (in the sense of data protection legislation).
- Personal data issued from the **research** and **participants management** are often not distinguished.
- **Data controllers** and **data processors** are not clearly identified (maybe it is in further documentation).
- **Anonymization methods** and **retention periods** are not specified.
- Inconsistent descriptions of how personal data will be handled throughout the project.
- Restricted access to data is often proposed without explaining whether it results from a risk assessment or from a default decision.

2. Legal and ethical aspects

Suzanna Farace Marazza

Common Privacy and Ethical Issues

Recommendations

- Clearly describe what personal (and sensitive, if any) data are collected, processed, stored, shared, and deleted, distinguishing between research data and participant management data.
- Identify data controllers and data processors.
- Specify anonymization techniques and retention periods.
- Ensure consistency between data reuse conditions, sharing practices, and licensing terms.
- Justify any access restrictions based on identified risks and demonstrate that data are shared **"as openly as possible, as restrictively as necessary**." The DMP should explain how potential harms were evaluated, what mitigation measures were considered, and why a more open level of access was deemed inappropriate.
- Good to store personal data on Swiss institutional servers (all 4 DMPs already do that though).

2. Legal and ethical aspects

Suzanna Farace Marazza

Few data protection definitions:

According to Art. 5 para. c. of the Federal Act on Data Protection, ***sensitive personal data*** is:

1. data relating to religious, philosophical, political or trade union-related views or activities,

2. data relating to health, the private sphere or affiliation to a race or ethnicity,

3. genetic data,

4. biometric data that uniquely identifies a natural person,

5. data relating to administrative and criminal proceedings or sanctions, or

6. data relating to social assistance measures.

According to Art. 5 para. j. of the Federal Act on Data Protection, the ***data controller*** is *a private* (physical or legal) *person who or federal body which, alone or jointly with others,* ***determines the purpose and the means*** *of processing personal data.*

According to Art. 5 para. k. of the Federal Act on Data Protection, the ***data processor*** is *a private* (physical or legal) *person or federal body that* ***processes personal data on behalf of the controller****.*

2. Legal and ethical aspects

Suzanna Farace Marazza

Common Copyright Issues

- Uncertainty about copyright ownership of materials created or provided by participants.
- Restrictions on data reuse may conflict with SNSF's open licensing requirements.
- The Creative Commons license (if any) to be applied is often not specified.
- Inconsistencies between stated reuse conditions and selected licenses.

Recommendations

- Clearly copyright ownership and usage rights.
- Justify any restrictions on data sharing and reuse.
- Specify the license that will apply to shared data and outputs.
- Inform participants about the implications of the chosen license.

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 1

Informed consent and opt out option

- Informed consent is obtained with written and oral study information, including specific data-usage agreements

- Participants are informed that they have the right to opt out of the experiment at any point in time resulting in the interruption of the data collection or, if data collection is completed, in the deletion of their data

Modular consent

Consent for data reuse Infromed consent for non-identifying data to be used, reused, or published in public scientific data repositories is given by an opt-in scheme in the informed consent form (i.e. a participant can decide which level of use, reuse, and sharing of their data they consent to). The options are the following:
- Data usage for study consent: - The use of the anonymised recordings

Files name contains participants' personal data

01_for_processing/ - script: 00_01_auto-rename.py - function: renames files based on the participant data and session metadata. Output recordings are formatted as sp{participantID}_sess{sessionID}_{age}_{sex}_{native_language code}_{experiment_language code}.wav.

Restrictive access to anonymised data

- Full anonymised version granted for research purposes only, after signing a data transfer agreement and agreeing to useage terms and conditions (restrictions for commercial use and unethical use of the data (e.g. creation of lie-detection systems))

- ✓ Informed consent
- ✓ Modular consent with opt out option

- Access to non-identifying data is restricted and subject to participant consent, although data that do not contain identifying information do not require consent for sharing.
- Anonymisation or pseudonymisation?
- File names contain participants' personal data

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 1

Harm minimisation

5.2 Necessary limitations to protect sensitive data

The project uses biometric data and collects sensitive data, both falling under the definition of *sensitive personal data* of the Swiss Confederation Federal Act on Data Protection (235.1 Federal Act of 25 September 2020 on Data Protection (Data Protection Act, FADP). This means that any potentially identifying recordings, annotations, or participant data will not be made public and will only be shared upon signing a data transfer agreement. Only

Special measures for audio recordings

> Audio recordings are released only after signing a data tranfer agreement and only if participant has opted in the sharing of their data to other research institutions

Anonymisation techniques

> - the use of text-to-speech technologies to create completely synthetic voices that resemble linguistically the original data (this excludes any potentially identifiable information in the recordings)

Not all data under the same license

Data will be published as open as possible as close as necessary under CC BY-NC license (CLARIN RES+PLAN+BY+NC+SA+*). Code will be published under the open source MIT license. > [!NOTE] > Not all data will be covered by the same license.

- ✓ Harm minimisation
- ✓ Special measures when sharing audio recordings
- ✓ Anonymisation techniques
- ✓ Not all data under the same license (that allows better sharing adaptations)
- ✓ MIT license for the code

- ○ It is unclear which data are covered by which license; this should be specified more clearly.

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 2

Appropriate selection and use of licenses

2.3 Copyright and Intellectual Property Right Issues

The ICE corpora are licensed, and the licenses also apply to material produced by our research project. New resources that are created will be made accessible by University of Zurich, the official distributor for the free ICE components (https://www.ice-corpora.uzh.ch/en.html) as far as the corpus creators allow. Derived language models do not allow the recreation of the original corpora and can thus be distributed open source, as traditionally done on websites such as huggingface (https://huggingface.co).

The main aim of the project is theory building and creation of annotated corpora (for licensed users) and open-access models (for everyone). We do not expect to create resources that can be commercially exploited. All three project partners will keep the right to use and extend all project data beyond the end of the project.

- Which licenses apply? What are the licensing terms for "licensed users" and "for everyone"? It is unclear which data are covered by which license; this should be specified more clearly.
- Is the reuse restriction for non-commercial purposes only compliant with the funder's requirements?

Important: Granting permission to the general public ("everyone") also requires a license (except for material that is not protected by copyright).

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 2

Free to use and public domain data

1.2 How data is collected, observed and generated

In addition to the existing data, further publicly available text sources and models are used, whenever possible from the public domain. In the case of texts to be annotated for training we will mainly use ICE, automatically annotated texts from open sources will be published. We will use statistical models and large language models, from the R repository (https://cran.r-project.org), huggingface (https://huggingface.co) and zenodo (https://zenodo.org).

No personal data collected: does this also apply to participant data?

2 Ethics, Legal and Security Issues

2.1 Ethical Issues

We do not record any data for this project. We use existing raw data. The ICE corpus, which is the most important text source, is available for research once licensed. Most components are available for free. University of Zurich is the official distributor for the free ICE components (https://www.ice-corpora.uzh.ch/en.html). All research partners (Zurich, Aachen, Erlangen) have licensed all ICE components.

- ✓ Prioritise the collection of public domain and open data, and clearly indicate their legal status.
- ✓ Minimise the collection of personal data and anonymise them as soon as possible.

- ○ It is unclear how participants' personal data are managed: are contact details collected, and if so, how are they anonymised?

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 3

Ethics and harm minimisation

2.1. How will ethical issues be addressed and handled?

The experimental data we will be collecting and using in this project is not bound to a confidential agreement because it does not contain sensitive information about participants. The experiments planned do not entail any kind of risks for participants and their personal information will not be recorded. Participants' dignity and health will not be impacted by the planned experiments. All the experiments will be presented to the Ethics Committee from the University of Geneva (CUREG2.0) in order to have their feedback and approval (https://cureg.unige.ch/en/).

Parents approval for underaged participants

For all types of experiments planned in this project, participants will receive an Information and Consent Form, that will be signed by participants (for the two older groups of adolescents, as well as for the adult control group) and their parents (for all adolescents).

Files name

The naming conventions of the files will be done as follows: type of data, type of experiment, name of the experimenter and the date (for example: Text_Elicitation_NameOfResearcher_25May2023). The

- ✓ Harm minimisation and Ethics Committee approval
- ✓ Parents approval for underaged participants

- ○ Avoid personal data in the files name
- ○ It is not clear which personal data is stored

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 3

Anonymisation technique

Participant gender and age will be recorded purely for the purposes of statistical analyses.

Inconsistency in data use purposes

The unanalyzed data will be available at the FORS database. Research peers who are affiliated with a university will be able to access and/or re-use the data for the purpose of analysis by signing a *user* contract (containing the obligations with respect to using, citing, and storing the data). Given that user needs to agree: 1. to use the data cited above and related instruments only for the scientific research and/or academic teaching outlined in my description in the present contract, and for no other purpose; 2. to use the data with respect to Swiss federal law and the applicable standard

What is meant by "registered"?

This project will not deal with sensitive experimental data. Participants' name, image, voice, or any other identifying information will not be registered.

- ✓ Data anonymisation in aggregated form
- ○ It is unclear which data will be shared for which purposes and there seems to be inconsistency in data sharing.

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 4

2 Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

The data is only used for scientific purposes. In compliance with the Federal Act on Data Protection – FAPD 235.1, the cantonal law on data protection LPrD, and the Federal Act on Research Involving Human Beings – (Human Research Act) HR 810.30, the PIs of the project and the postdoctoral researcher submitted the research protocol, covering the project as a whole, to the *Commission d'éthique de la recherche* of the *Faculté des Lettres* (Université de Lausanne) for clearance. According to our research, a separate formal ethical clearance procedure was not required for fieldwork conducted in India and Sri Lanka. Ethical clearance for the Switzerland fieldwork was obtained through the *Commission d'éthique de la recherche* of the *Faculté des Lettres* at the University of Lausanne by the PI and the postdoctoral researcher prior to the commencement of fieldwork. For fieldwork in India and Sri Lanka, the collaborating PI obtained an

ethics clearance certificate from the Indian Institute of Technology Jodhpur.

- ✓ Authorisation from Ethics Committee to share personal data abroad
- ✓ Good coordination between research teams supports data protection, security, and participant safety.

- ○ The Federal Act on Research Involving Human Beings does not apply in this case, as Art. 2 states that it applies on "research concerning human diseases and concerning the structure and function of the human body"

2. Legal and ethical aspects

Suzanna Farace Marazza

Sample 4

Lack of clear licensing terms

- All users of shared datasets are required to comply with the applicable Creative Commons licence.

Personal data access and retention

- Full access to non-de-identified data is restricted to project team members and authorised student assistants working on the project. Access to linguistic data and de-identified metadata may in future be granted to researchers upon request and verification of research identity, in accordance with participant consent agreements and applicable data-sharing
- For presentations and publications, no personal information that could lead to the identification of participants is disclosed. Instead, participants are referred to only through unique participant codes linked to the dataset;

Each participant is associated with a unique code to de-identify the participant´s information. Personal identifiable information, including participant names, place of birth, etc. is stored separately from the research dataset and is accessible only to authorised project members.

Personal information, including names, locations, and other identifying details, was de-identified in the transcribed data and excluded from publicly shared datasets. Access to non-de-identified data is restricted to authorised project team members in order to ensure data protection and participant confidentiality.

- ✓ It is good to strongly restrict access to datasets containing personal data and to share anonymised datasets

- ○ It is unclear whichCreative Commonslicense(s) will apply towhich dataset(s); theapplicable licensingterms should be clearlyidentified and justified
- ○ Retention, deletion, and anonymisation periods for personal data are not clearly defined