Plan Overview

A Data Management Plan created using DMPTuuli

Title: Snabbmeddelanden på flera språk: WhatsApp i finlandssvensk digital kommunikation

Creator: Ines Fröjdö

Affiliation: Hanken School of Economics

Template: Hanken's DMP template

Project abstract:

This project will collect a multi-modal text corpus with the objective to map current trends of loanwords and code-switches occurring in the instant messages of young adults using Finnish-Swedish in a multilingual environment.

ID: 20695

Start date: 01-08-2022

End date: 31-07-2025

Last modified: 03-02-2024

Grant number / URL: SKF application number 178990

Snabbmeddelanden på flera språk: WhatsApp i finlandssvensk digital kommunikation

General information

- **Give a brief description of your research such as its name, research members, funding information and project number. For example, if it is an official research project at Hanken, what is Hanken's project number?**
- **What is the version of this DMP? Is it an initial, detailed or final version? Also specify here the date when this version of the DMP is written.**

Detailed DMP, updated 12.12.2023
Research project at the Centre of Language and Business Communication at Hanken School of Economics. The Principal Investigator is Martti Mäkinen and the research assistants are Leyla Shojaeifard and Ines Fröjdö.
The project will produce a corpus of Multilingual / Multi-Modal WhatsApp discussions at Hanken or MMWAH. Funding for two years of research has been granted from Svenska Kulturfonden.

1. Data description

1.1 Give a brief description of your data. Answer the questions:

- **How *new data* will be collected or produced in the project?**
- **How *existing data* will be reused?**
- **What *kinds* of data will be collected, produced or reused?**
- **In which *file formats* will the data be in?**
- **And estimate the *data size* if you know.**

A new dataset will be collected and processed during the run of the MMWAH project. Data will include chat data, such as participant names, pictures, time stamps, and message contents. The donations will likely contain strong direct and indirect personal identifiers. Metadata about each participant will be collected through a questionnaire sent out to the participants. It will enquire their demographic information and linguistic background.

Data type	Source	Format	Sensitive?	Size
Chat data	collected	.txt /.jpg /.mp3 / .xml	yes	TBA
Questionnaire	collected	Webropol / .xlsx	yes	TBA

All data collected by email, including videos and recordings, will be transferred without delay into and stored in a secure, shared folder in IT-systems provided by Hanken, and accessible only by the research group members. The files on the project email account will then be deleted.
Preference will be given to open and standard data formats to facilitate sharing and long-term reuse of the data.

1.2 Answer the question how the consistency and quality of data will be controlled. Careful documentation of the procedures of data collection is the primary measure to ensure the integrity and quality of the data.

To ensure the quality and consistency of the data the dates of data collection, retrieval, and changes will be recorded, making all data-related actions traceable and repeatable. Version control will be implemented once the data has been manually psedonymised. Data protection measures such as minimisation, pseudonymisation, and anonymisation will not affect data quality, as the minimisation is done deliberately and the pseudonymisation methods will help retain demographic data. In all conversions, maintaining the original information content will be ensured by allowing the participants to review and pseudonymise their data.

2. Ethical and legal compliance

2.1 What legal and ethical issues are related to your data management, for example, in handling personal data, protecting the identity and rights of participants, gaining consent for data sharing and publishing?

Our research contains personal data that is gathered in the form of group chats submitted to us by one of the chat participants. All the research members bear the responsibility for ethical and moral concerns and decisions involved in the research and during the interaction between the researcher and research participants in accordance with relevant legislation, research integrity, and good science practices to maintain high ethical standards and comply with data protection laws.
The informed consent and privacy notice that research participants are provided to read and accept will be kept on file and made available upon request. The donors consent to their data being used, stored and published as a part of the donation process.

2.2 Agreements on data ownership and other intellectual property rights such as secondary data usage copyright permissions and open data licenses need to be concluded before commencing any actual research activities. How will you manage the rights of the data you (re)use, produce, share, and publish?

The principal investigator is responsible for concluding contracts on authorship, data ownership, data sharing, and user rights, which will be agreed on with all researchers prior to the start of actual research. The data ownership agreement describes who owns the data, and whether and what rights will be transferred. Copyright and intellectual property rights will also be secured before any data is made public.
The data will be published for reuse under the Creative Commons license CC BY-NC-SA 4.0 and made available for use on Open Access corpus repositories, such as Korp, Kielipankki and Språkbanken. The project will be given the permission by research participants to publish their data in an anonymised form.

3. Documentation and metadata

Describe here what kind of metadata standards, README files or other documentation you will use to help others understand and use your data, and make your data FAIR.

The research project endeavours to ensure the findability and citability of the research data in line with the FAIR data principles, while sees to that the degree of data openness and sharing is ethically and legally justifiable.

The research project will make sure that properly documented metadata of research data is published by using the Fairdata Qvain metadata tool. Qvain is part of the Fairdata services to support research data to go FAIR. The services are offered by the Ministry of Education and Culture and produced by CSC.

To make our data easily findable, the project will archive the digital data in IDA, Aila, and/or Zenodo and use descriptive metadata as required and provided by Kielipankki and Språkbanken repositories.

The open research data and associated metadata will be assigned unique DOI (digital object identifiers) enabling (meta)data uniquely identifiable, and thus accessible and referenceable.

Research datasets are registered in Hanken's research database Haris with the persistent identifiers for the (meta)data.

To make our data openly accessible, tools, software, and components used in the project will be available as open source as far as privacy allows.

To make the data accessible and interoperable, used formats will be based on open standards. When depositing data in the above repository, the project will ensure that the research data is migrated to new formats, platforms, and storage media as required by good open science practice to enable data sharing, reuse, and interoperability between researchers, institutions, organisations, and countries.

Files and folders will be versioned and structured by using a name convention consisting of project name, dataset name, and version information. Search keywords and subject headings from the KOKO Ontology (integrated in the Qvain metadata tool) will be provided to optimize data reuse possibilities.

For the data to be reusable, the Creative Commons license CC-BY-NC-SA 4.0 will be used for the project's outputs and open research data, free of charge for any user and without any embargo period, to ensure that they are shared with minimal restrictions, aside from attribution to the authors or creators.

Before submitting and depositing any part of the corpus created in the above repositories, all direct identifiers will be removed, indirect identifiers removed or categorized, and all sensitive personal information deleted to ensure that the data is properly and irreversibly anonymized. The donated raw data will not be submitted for archival.

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will they be backed up? How will you share the data securely with your research partners?

During the active research period, the data will be stored in and shared through the information system provided and maintained by Hanken. Hanken-provided systems do automatic backups. Data are retrievable in case of human error or data corruption. In addition, manual backups of master data files will be taken regularly and always before any major file-format or data conversions.

4.2 Who will have access to the data during the research? Who will be responsible for access control? And how will secured access be controlled?

Right to access the data and data usage are controlled by the PI. The PI completes the list of users and all rights granted, and a procedure for withdrawing rights. Technical access control is provided by IT-services of Hanken. Data will be available to all research members of the project by using the 'Specific people'-option in Hanken's OneDrive portal, to keep control over who can be the authorized users.

Access control will be in line with the level of confidentiality involved. Data with direct identifiers, contact information, sensitive personal data, and confidential data will be protected with adequate, additional appropriate safeguard measures such as encryption and strict access control, and will not be sent between research team members by email – not even Hanken's email system.

5. Opening, publishing and archiving the data after the research project

5.1 What part of the data can be made openly available or published? Where and when will the (meta)data be made available?

Data in the form of a corpus will be available and cited in peer-reviewed scientific journals. The project ensures that all the publications will be immediately openly accessible. Legitimate copies of the publications will be uploaded and preserved in Hanken's institutional repository, DHanken.

The project will be given the permission by research participants to publish the data in an anonymised form in language data repositories for the purpose of allowing other researchers as well as scientific publication outlets to conduct further scientific analyses on the data. The originally donated data as well as the data collected from the demographic questionnaire are not anonymised and will not be submitted for archival.

The digital data and the associated metadata will be archived in the data storage repository:

- Zenodo by the OpenAIRE project and CERN,
- IDA - Research data storage and archival service part of the Fairdata services by the Ministry and CSC,

5.2 Will data with long-term value be preserved? If yes, where and for how long?

The research team will discuss and consider our data's long-term value and preservation for more than 25 years.

Costs related to long-term preservation will be calculated and covered by the grant received from Svenska Kulturfonden.

6. Data management responsibilities and resources

Who will be responsible for specific tasks of data management during the research project life cycle? Also estimate the resources (e.g., financial, time and effort) needed for the data management tasks.

The project PI is responsible for the initial planning and execution of data management procedures. The researchers Leyla Shojaeifard and Ines Fröjdö working for the project are responsible for the tasks of data collection, data quality, data storage and backup, metadata production, data archival and publishing. Hanken's research support unit will provide guidance for data management/stewardship and management of IPRs.

The research project team members have allocated time and budget to complete the data management tasks and cover relevant costs. Archival in the repositories is free of charge. The data

management tasks altogether will take around 30 months during the 3-year's project.