Plan Overview

A Data Management Plan created using DMPTuuli

Title: Designing Inclusive and Trustworthy Digital Public Services for Migrants in Finland (Trust-M)

Creator: Nitin Sawhney

Principal Investigator: Aaro Tupasela, Thomas Olsson, Suvi Sankari, Tom Bäckström, Johanna Ylipulli, Teemu Haapalehto

Affiliation: Aalto University

Funder: The Research Council of Finland (former The Academy of Finland)

Template: Academy of Finland data management plan guidelines (2021-2023)

ORCID iD: 0000-0003-1512-7533

ORCID iD: 0000-0002-1106-2544

ORCID iD: 0000-0002-1410-6602

ORCID iD: 0000-0002-5590-2349

ORCID iD: 0000-0002-7537-4774

Project abstract:

The Trust-M project aims to improve the integration of migrants in Finland by devising hybrid and trustworthy digital services based on conversational AI. This can strengthen social cohesion, resilience of the labor market, and economic vibrance in Finnish Society. To this end, we seek to understand how socially and culturally constructed notions of trust as well as human rights are incorporated in present-day digital public services and perceived among migrant communities. Finnish public services may not always be accessible, inclusive or trustworthy for all migrants. Using conversational interaction and hybrid service design, Trust-M will explore, create and pilot innovative AI-assisted services. The City of Espoo, as the main interaction partner, will ensure their practical relevance to migrant services and promote societal impact. Trust-M offers key insights for how to meaningfully empower and effectively integrate diverse migrants through digital public services in municipalities.

ID: 19977

Start date: 01-10-2022

End date: 30-09-2025

Last modified: 14-06-2023

Grant number / URL: 353509

Designing Inclusive and Trustworthy Digital Public Services for Migrants in Finland (Trust-M)

1. General description of data

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in? Additionally, give a rough estimate of the size of the data produced/collected.

Trust-M produces, analyses, and manages data in the following categories:

Data Type	File Format	Size	Sensitive	Source	Storage location
1. Interview audio (digital audio recorder)	.mp3 / wav	≈25MB each	Yes	Data Collected	Aalto secure server
2. Interview transcripts	.txt / .xls / .docx / .atlproj	100kB - 1MB each	Yes	Data Produced	Aalto secure server
3. Observational fieldnotes	.txt / .docx	100kB - 1MB each	Yes	Data Collected	Respective university devices
4. Online surveys (Eg. REDCap)	.csv / .txt / .docx	≈10MB total	Yes	Data Collected	Aalto secure server
5. Photographs	.jpg / .png / .gif	1MB - 10MB	Yes	Data Collected	Aalto secure server
6. Video Recordings (no cloud upload)	.mov / .mp4	10MB - 100MB	Yes	Data Collected	Aalto secure server
7. Public documents and reports	.pdf / .docx / .html	100kB - 10MB	No	Data Reused	Respective university devices
8. Confidential or non-public documents and reports	.pdf / .docx / .html	100kB - 10MB	Yes	Data Reused	Aalto secure server
9. Design artifacts (mockups, sketches, storyboards, blueprints etc.)	.png / .img / .gif / .pdf	100kB - 10MB	No	Data Produced	Aalto secure server/Respective university devices
10. Chatbot training data	.xlsx / .csv / .txt / .json	0.5MB - 10MB	No	Data Reused	Respective university devices
11. Speech, language and acoustic analysis and training data	.mp3 / .wav / .txt / .csv	10MB - 10GB	No	Data Collected and Reused	Respective university devices
12. Source code	.py / .txt	100kB - 10 MB	No	Data Produced and Reused	Respective university devices

1.2 How will the consistency and quality of data be controlled?

In general, we will use systematic protocols specific to different data and research methods, and these protocols will be periodically reviewed by the PIs. The journals and other outlets publishing data will perform their own quality checks regarding scientific quality and research integrity, as part of the peer review process. When necessary, the project personnel will attend their home institutions' training sessions regarding data management and research integrity. All digital material is regularly backed up by the universities' file management systems. All physical material is stored in a safe location in university premises with limited access and security surveillance systems.

The following outlines more specific consistency and quality measures related to each of the data types presented in 1.1.

1. The interviews and focus groups are based on carefully planned interview scripts, ranging from unstructured thematic interviews to semi-structured interviews. The work package leaders are in charge of consistency regarding substance and the quality of data management. To ensure good audio quality, we will use multiple audio recording devices in the planned interviews and co-design sessions.

2. Interview data will be initially stored as audio files and further transcribed into textual transcripts using professional services (eg Tutkimustie), and the transcripts serve as the primary data in the analysis phase. The contracts with the service providers will strictly adhere to the personal privacy and confidentiality agreements of the respective universities conducting the research. The quality of the transcripts will be controlled by the researcher who conducted the interviews and compared with audio files if necessary. Permits and informed consent forms will be scanned or photographed and exported into PDF files.

3. Observational field notes follow the same guidelines as data types 1 & 2.

4. Online surveys follow largely the same guidelines as interviews. Additionally, we will use professional language proofing services and run small-scale pilot surveys to ensure the understandability of the survey questions.

5–6. Concerning photographs and video recordings, we will ensure that the photographer has the necessary skills to capture technically high-quality and semantically descriptive material, without compromising the research subjects' privacy.

7–8. For external documents, we will ensure that we have the most recent versions and that we correctly interpret the purpose and intended audience of each document.

9. As the sketches and other output from co-design sessions tend to vary in terms of detail and readability, will ensure that the key ideas and perspectives behind each idea are well documented as soon as possible after the sessions. In addition to taking photos of the produced material (e.g., post-it notes, sketches on paper), we take notes that aim to capture the general discussion in the sessions. More detailed and hi-fidelity design proposals may be managed and stored in external design tools, such as Figma or Miro, however ensuring that they do not disclose any personal information.

10–11. The training data has undergone quality control by the party that generated it.

12. We will follow high-integrity practices of documenting the code as well as describing the produced solutions when publishing them in GitHub and the like.

2. Ethical and legal compliance

2.1 What legal issues are related to your data management? (For example, GDPR and other legislation affecting data processing.)

The consortium is committed to following the guidelines issued by the Finnish Advisory Board on Research Integrity ("*Responsible conduct of research and procedures for handling allegations of misconduct in Finland*") on good scientific practice, how to handle violations against it, as well as valid legislation. We are also following the European Code of Conduct for Research Integrity by ALLEA.

Audio and video data cannot be fully anonymized and therefore we need to retain identity information of participants so that they can withdraw consent if required. Withdrawal requires proof of identity. Collected video and audio will be retained for 2 years past the end of the project and then deleted.

All participation in surveys, interviews, focus groups, and workshops will be voluntary and individual participants will be asked to provide informed consent, covering both research and data sharing. All participants will be de-identified in the data analysis process when possible. The participants will also be given a privacy notice, which explains the data collection policies and protocols, and subsequent analysis and archiving for the project in accordance with GDPR guidelines. All participants will be legally competent adults; in cases where youth are interviewed, consent from their guardians will be requested as well.

Our qualitative research with participants does not intend to address sensitive topics or personal information regarding their trade union membership, data concerning health, sexual orientation or activity, or genetic and biometric data for identifying the person. However, participants' racial or ethnic origin, religion or philosophical beliefs may be addressed in some instances, in which case we will strive to handle such sensitive topics with care while ensuring we inform participants if they are willing to share such information voluntarily and that it would not be used in any way that would be harmful to them.

For any data collected will follow each university's data policies for all identifiable human data collected:
https://www.aalto.fi/en/services/aalto-university-data-protection-policy

https://www.tuni.fi/en/data-protection/data-protection-policy

https://www.helsinki.fi/en/about-us/data-protection

Processing of all personal data will be handled according to Aalto's practices and guidelines:
https://www.aalto.fi/en/services/how-to-handle-personal-data-in-research

Our compliance with the requirements stipulated by the EU General Data Protection Regulation (GDPR), national data protection legislation and other legislation that relates to the processing of personal data will be ensured as per the data policy. The collected data will only be stored on systems meeting Aalto University, Tampere University, and University of Helsinki guidelines for personal data storage. Ethical pre-approvals are not generally needed.

For research undertaken with human participants, ethical approval will be requested by Aalto Research Ethics Committee and/or Ethical Committees of Tampere University or University of Helsinki, depending on which party takes the lead in data gathering.

2.2 How will you manage the rights of the data you use, produce and share?

As a principal rule, the Consortium Universities will own the data produced by research personnel working on the project. The Trust-M project manages the rights for the research through a *Research Cooperation Agreement* signed by all participating universities. The agreement will be finalized before the end of 2022, and it will follow the general guidelines provided by Aalto University.

As several work packages in Trust-M will work across Universities and use shared datasets, the *Research Cooperation Agreement* will cover aspects of cross-institutional data processing, transfer and joint controller issues. At this point, we do not foresee international collaboration that would call for moving any personal data outside EU. However, as international networks mobilize, the consortium leadership will be responsible of making sure necessary GDPR requirements are followed.

As the project progresses any anonymized data collected that would be useful to share in the public interest may be made available under theAalto Open Science and Research Policy. In order to comply with the European Open Data Directive, the above selected data will be made openly under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

For the reused data sets (types 7, 8, 10, 11 & 12), we will follow the use agreements made with the data providers with respect to data handling and dissemination.

The rights for the anonymized data collected for the project will be defined in the cooperation agreement between the research consortia partners. All authors of the software will agree on both the open release and licensing of the software.

3. Documentation and metadata

3. How will you document your data in order to make the data findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?

The consortium is committed to following the FAIR data principles in order to make our data findable, accessible, interoperable, and reusable. Such a strategy will promote joint use of data during the research and ensure the interpretation of data and their re-usability after the project. During the project kick-off meeting, under the guidance of the PI, the consortium will define the common practices on how to implement the FAIR principles in detail. The data generated during the project will be converted to an electronic format. The folder structure will separate the different types of data. The filenames will contain all descriptive information as defined below for easy retrieval afterwards. Datasets will be documented with descriptive metadata which ensures the understandability and findability of the data in the future. If needed, README files will be created for datasets to ascertain their re-usability, reading, and interpretation.

Data will be made:

- Findable by giving the metadata and published data permanent identifiers.
- Accessible by giving controlled access to data when allowed by the ethical guidelines and GDPR
- Interoperable by using standard file formats used in the field.
- Reusable by making rich metadata to accompany the datasets.

The data will be README-documented using the FAIR principles when applicable, i.e., conforms to ethics and privacy, consent of the included subjects, and ethical clearance. Metadata will be saved with the following information:

- Name of the dataset
- Description of the dataset
- The creators of the dataset
- Creation date
- Size and format
- Access type
- Version information (draft, final, published etc.)
- Information on each variable of the data

The naming of data files will follow a protocol agreed upon at the beginning of the project, e.g. (TrustM-WSxxxxDD), where W=WP number, S=study number, xxxx=participant code, from 0001, DD=data type.

4. Storage and backup during the research project

4.1 Where will your data be stored, and how will the data be backed up?

We will use institutional data storage systems that take into account the appropriate security level of the data and the needs born from our multidisciplinary collaboration. Each consortium party will use its IT services and secure cloud storage services to store data during the project. During the project, data will be stored primarily on the institutional secured servers, which are supported by a snapshot feature and regular backups that make file versions automatically recover from unwanted deletions. For temporary purposes, such as taking notes during an interview, the data will be stored on researchers' laptop computers and later transferred to cloud storage. All the laptops utilized in the project are administered by consortium universities' IT support. They include a secure file transfer over the network with a VPN solution. The IT services provide regular backup copies of the data content of servers under its administration, including personal cloud folders.

If necessary, computational training data (10–11) may be stored on the servers of the Finnish IT Center for Science.

Aalto University's internal secure data storage service (Secwork) would be used as strictly access controlled and monitored storage space for sensitive data. Only authorised researchers in the research consortium would be provided user accounts to access Secwork to facilitate secure sharing of sensitive research data as needed.

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?

For each WP, their leader is responsible for granting access to the collected personal data (e.g. types 1, 2, 4, 5 & 6) in accordance to their institution's policy. When data are shared between groups from different institutions, whichever institution has stricter policies, those policies are followed.

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 data, or its metadata, be made available?

Trust-M aims to make as much of the data available as possible, mindful of copyright and privacy restrictions. Relevant data may be published as part

of scientific articles or in separate repositories at the same time as the articles are published.

Software code will be shared as GitHub repositories with suitable licenses (e.g., GNU GPL, MIT, or Apache 2). The deposited data and code will be supplied with the required standard metadata to ensure the reusability and ensuring reproducibility of results in publications. Possible data related to the published materials will be made available in Zenodo or similar repositories under the Creative Commons license CC BY. The repositories also provide persistent identifiers (e.g., DOI, URN) to promote citations. The selected published data will be accompanied by metadata as described in Section 3. Personal data that cannot be efficiently anonymized (and thus cannot be published) will have its metadata made public through the institutional research information systems (e.g. in Aalto case, the ACRIS). We will use suitable paper pre-print services, e.g., the ArXiv.org e-Print archive, whenever needed.

5.2 Where will data with long-term value be archived, and for how long?

The codes and scripts developed in the project, as well as the shareable data, will be made available on public code and data repositories, such as Github and Zenodo, which will guarantee their long-term preservation. Forms (informed consent form etc.) in paper format will be destroyed securely after revision time. The digital material will be archived as long as the research data will be stored. The storage period will be defined during the project. We will contemplate the possibility of archiving data with recognized long-term value in the national Fairdata PAS service.

Social scientific data (anonymized transcriptions) will be archived in Finnish Social Science Data Archive:https://www.fsd.tuni.fi/en/

6. Data management responsibilities and resources

6.1 Who (for example role, position, and institution) will be responsible for data management?

Each researcher collecting or storing data for the project will be responsible for the management of the data they have stored. That is, data management is done by each researcher as an integral part of data collection and analysis. When data is stored in a shared space such as a group folder, the PI, or a person they have authorized, will have admin rights to that shared space.

PIs will generally be responsible for ensuring that data management practices are followed. For data in shared spaces, at the end of the project, the administration rights will be assigned to a suitable owner, most likely the responsible PI, and any outstanding issues related to size and/or cost of the folder will be coordinated with the relevant university's data support.

PIs are also responsible for ensuring that the DMP is regularly reviewed and revised if necessary.

6.2 What resources will be required for your data management procedures to ensure that the data can be opened and preserved according to FAIR principles (Findable, Accessible, Interoperable, Re-usable)?

As stated above, the data management procedures are integrated to the practical conduct of research in different WP's. The data management will rely on resources offered by each consortium university, CSC and public open services that promote open science (e.g., Open Science Framework and Zenodo).

Both Aalto and CSC provide guidelines as to how to produce reusable, interoperable implementations and analyses. Aalto University's version control system (https://version.aalto.fi), as well as one provided by Github, are at our disposal to make our code and datasets accessible and findable.