Plan Overview

A Data Management Plan created using DMPTuuli

Title: Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo)

Creator: Ilona Pikkanen

Principal Investigator: Ilona Pikkanen, Mikko Koho, Jouni Tuominen

Affiliation: Other

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

Template: Academy of Finland data management plan guidelines

ORCID iD: 0000-0001-9435-7163

ORCID iD: 0000-0002-7373-9338

ORCID iD: 0000-0003-4789-5676

Project abstract:

This multidisciplinary consortium proposal seeks to create an unprecedented research resource for the Finnish humanities. It integrates letter catalogue metadata from the cultural heritage (CH) institutions in Finland into a single reconciled Linked Open Data infrastructure and publishes analysis tools on an open semantic portal, where a researcher can discern at a glance a whole web of connections with her object of study, the available letter collections and their keepers. The project enables us to conduct empirical, bottom-up case studies on epistolary culture and social networks in the Grand Duchy of Finland (1809–1917), ask ambitious research questions in the field of computer science and make currently scattered, heterogeneous epistolary metadata interoperable and available. The timeliness of the proposal is optimal because it can build on the results and experiences accumulated in the COST Action Reassembling the Republic of Letters (RRL), 1500–1800 (2014–18). The proposed consortium will work with a dataset of hundreds of thousands of letters, which is temporally and geographically more restricted than the one used at RRL but much more representative regarding social aspects. The computational teams, with strong background in multidisciplinary Digital Humanities research, will research topics around the metadata integration and harmonisation, as well as study and develop methods and tooling for data analysis on correspondence metadata, with the goal of providing answers to humanistic research questions. The harmonised data provides us with a quantitative understanding of Finnish epistolary culture and its changes over time. Questions related to social stratification and gender are analytically more complex, but, when confronted with our bottom-up analyses, they can be posed and answered in new and more precise ways, from a novel vantage point of Finnish 19th-century society and its networks of written communication. The consortium brings together three teams from the University of Helsinki, Aalto University, and Finnish Literature Society, with complementary, multidisciplinary expertise in the central scholarly approaches of the project, its key methodologies and the construction and publication of open semantic portals. The project can only be carried out in close collaboration with key CH institutions, and we have ensured their full support.

ID: 16630

Start date: 01-09-2021

End date: 31-08-2025

Last modified: 22-06-2021

Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo)

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.

The main providers of letter catalogue metadata harmonized and analyzed in the project are the National Archives, the National Library, the Finnish National Gallery, the Finnish Literature Society, and the Swedish Literature Society (smaller datasets are e.g. at the Theatre Museum and the Labour Archives). Further collections will be surveyed in the first phase of the project. The cultural heritage (CH) institutions do not know how much correspondence (both between individuals and to public bodies and institutions) there is in their collections, and how large a part of it has been catalogued. The preliminary surveys have established the following rough estimates: metadata of more than 260,000 letters is readily available in some digital structured format, i.e. in databases or XML files. Full-text content is available of at least 9071 letters.

Overall, the data will be heterogeneous, varying in formats, quality, and modality. The letter catalogue metadata is mostly simple, containing names of senders and recipients, and often dates and places of writing, while some catalogues also contain rich metadata based on the letter contents, such as language and referenced people, places, events, and artworks. The reading or reusing of the data does not require uncommon software.

The project produces a Linked Dataset of the letter catalogue metadata by converting the data acquired from CH institutions. The dataset is in Linked Data (RDF) format and queryable via SPARQL endpoint, with CSV, XML, and JSON output formats. The size of the dataset is unknown at this point.

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

A central challenge regarding epistolary metadata is the semantic disambiguation of person references, i.e. the task of removing uncertainty of meaning from possibly ambiguous textual representations or structured metadata. This requires that the used data model, user interfaces, and data analyses are able to deal with both discrete and continuous time. We will use standardized protocols for exact recording of uncertainty and ambiguity, which enables filtering out such cases when needed.

In conversion processes the original data is maintained in order to ensure the transparency and reproducibility of the process. The produced Linked Dataset and the software code for generating it will be stored on a version control platform, which enables the tracking of the changes made to the data and code during the project by utilizing the version history.

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 letter-catalogue metadata will be harvested until 1917 since the EU's General Data Protection Regulation (GDPR) does not apply to information relating to a deceased person. We will, however, scrutinise carefully that the potential information regarding special categories of personal data will be processed ethically and taking into account, when necessary, what is regulated on personal data of living persons. The possible problems regarding metadata of collections with restrictions on use will be carefully assessed (although preliminary discussions with the National Archives indicate that the epistolary metadata can be utilised even in these cases).The data collection, management, and storage will follow Finnish legislation (concerning data protection and copyright).

The consortium is committed to following theguidelines issued by the Finnish Advisory Board on Research Integrity 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

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

There are no copyrights, licenses or other restrictions that prevent us from using or sharing the data. However, this will be further specified in the agreements the consortium will make with the data providers (CH-organisations), when necessary. The overarching objective of the consortium is to produce an integrated dataset that will be made available to the research community via an open data service and semantic portal. The consortium parties will thus collaborate closely in the processing of the data: each consortium party will work with the same datasets throughout the project. These data and results will be jointly owned by University of Helsinki, Aalto University, and researchers from SKS; researchers working at the University of Helsinki and Aalto University will give an Undertaking on Transfer of Rights as requested by the university which will transfer his or her rights to all the results achieved during the research to the University of Helsinki and Aalto University, respectively.

The aim is to publish the data as openly as possible (using e.g. licenses CC0, CC BY). There are no possibilities of commercial exploitation of the data.

3. Documentation and metadata

3.1 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 produced Linked Dataset and its metadata will be stored using W3C's open, machine-processable RDF standard, using established data models (e.g. Dublin Core, CIDOC CRM), vocabularies, and best practices of the domain, identified during the project. The data will be openly accessible and reusable, both as individual data items and as the entire data collection, based on Linked Data standards (RDF, SPARQL) and best practices, using persistent identifiers (HTTP URI). The metadata documentation of the data collection will be based on data cataloging standards, including DCAT and VoID. The produced dataset will be accompanied with a documentation web page describing the data, its data model, how the data was collected, how it can be used via Linked Data standards with example queries. The archived dataset will include a README file describing the dataset.

4. Storage and backup during the research project

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

The source datasets and the produced harmonized Linked Dataset will be stored in Aalto University's and/or University of Helsinki's GitLab version control platform (http://version.aalto.fi, http://version.helsinki.fi), which contents are backed up by Aalto University and University of Helsinki IT services, respectively. The storage services of CSC (e.g. Allas object storage) can be used in

addition, as needed.

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

All consortium institutions hold high security standards, and data is stored in password-protected environments that ensure secured access to the data only to authorised users involved in the research project. During data analysis, the data will be accessible only to the researchers involved in the project. In the publication phase (open data service and semantic portal) special attention will be paid to the data that may have restrictions of use in the collections of the archives (although our preliminary discussions with the CH institutions imply that even in these cases the metadata is openly publishable).
Access to the collected source datasets and the developed Linked Dataset will be controlled by the project personnel. Access to the data in the GitLab version control platform is restricted to the project personnel and other members of the research groups the project personnel work in (e.g., Semantic Computing Research Group (SeCo) at the Aalto University). Access to the CSC storage services is restricted to the project personnel.

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?

The letter metadata collections of the cultural heritage institutions are available with an open licence, mostly Creative Commons 0 (Public Domain). The project will publish the harmonised metadata with an open licence which will enable its later use in CH institutions. The data will be findable, accessible, interoperable, and reusable satisfying the FAIR principles. The data will be published on the Linked Data Finland (LDF.fi) data publishing platform run by Semantic Computing Research Group in Aalto University. The LDF.fi service is hosted on the cloud computing service cPouta / container cloud Rahti run by CSC – IT Center for Science. The data can be downloaded as individual data items or as the entire data collection, or parts of it can be queried. The Linked Open Data service is detached from the application layer, which makes the system reusable for everybody. The harmonized dataset will be published also on Zenodo. In addition, software solutions and DH tools of the application layer will be published open source under MIT Licence, maximising reuse also in a commercial setting.

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

Suitable options for long-term preservation (including CSC DIPI infrastructure) of the produced research data will be investigated during the project. The storage period will be decided during the project.

6. Data management responsibilities and resources

6.1 Who (for example role, position, and institution) will be responsible for data management (i.e., the data steward)?

The consortium parties will share the responsibility for data management during the research project life cycle. The more particular responsibilities are as follows. Team 1 (the Finnish Literature Society) will employ a data manager, who will be in charge of WP1 (collection and curation of epistolary metadata from the CH institutions). Team 2 (Aalto) is in charge of integrating and harmonizing data and of data disambiguation and enrichment. Team 3 (Heldig) is in charge of data analysis method development and the semantic portal development. Research support services (Aalto) and data support services (University of Helsinki) will be consulted when needed to ensure that data management best practices are followed.

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)?

Research integrity and open scholarship lie at the heart of this data-driven, 4-year consortium project. Most of the resources of the project are allocated to data management. All the work will be carried out according to FAIR principles. The storage, preservation, and sharing of the data does not incur expenses.The computational resources are provided by the partner institutions and Ministry of Education and Culture (CSC – IT Center for Science). As part of the project, the data will be produced in formats that can be imported into research data storage and preservation services.