Searching, exploring, and analyzing historical letters and the underlying networks: LetterSampo Finland (1809–1917) data service and semantic portal*

Eero Hyvönen[1,2,*], Petri Leskinen[1], Henna Poikkimäki[1], Heikki Rantala[1], Jouni Tuominen[2,3,*], Senka Drobac[2], Ossi Koho[2], Ilona Pikkanen[4,*] and Hanna-Leena Paloposki[4]

[1]*Aalto University, Department of Computer Science*
[2]*University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)*
[3]*University of Helsinki, Helsinki Institute for Humanities and Social Sciences (HSSH)*
[4]*Finnish Literature Society*

Abstract

Epistolary data (data related to letters) is by nature stored in geographically distributed archives and collections, as letters are exchanged between different people and places. To get a global view and analyze correspondences, the archival collections, and the underlying egocentric and social networks, data from the separate data silos in different cultural heritage (CH) organizations have to be aggregated, harmonized, and published as a global data service with APIs for Digital Humanities research and application development. This paper presents an overview of the system *LetterSampo Finland (1800–1917)* consisting of a Linked Open Data (LOD) service and a semantic portal designed for these purposes. The LOD service contains extensive metadata on one million letters sent or received in the Grand Duchy of Finland during 1809–1917, aggregated from various Finnish CH organizations, harmonized by using a shared ontological data model and vocabularies, and published as a LOD service with a SPARQL endpoint and data dumps under an open license. Based on the so-called Sampo model and Sampo-UI framework, a new semantic portal has been created on top of the LOD service. This portal can be used for searching, exploring, and analyzing letters, letter collections, and networks within these correspondences.

Keywords

linked data, epistolary data, semantic portal, data analysis, network analysis,

1. Introduction

Letters are an important source of data for historical research, biography, and prosopography. Letters have been in a central role for the development of scientific thinking: During the Age of Enlightenment it became possible for people to send and receive letters across Europe and beyond, based on a revolution in postal services. This opportunity resulted into the so-called

Digital Humanities in Nordic and Baltic Countries, Tartu, Estonia, March 5–7, 2025
*Abstract proposal for a long paper.
*Corresponding author.
eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/eahyvone (E. Hyvönen)
0000-0003-1695-84 (E. Hyvönen)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

Republic of Letters (RofL) (Respublica litteraria), a cross-national collaborative communication network that formed a basis for modern European scientific thinking, values, and institutions in Early Modern times 1400–1800 (Hotson and Wallnig 2019; Miert 2016). Sending letters is a phenomenon that is in many ways analogous to many means of communication using the Internet, email, social media, and the World Wide Web (WWW) since the 1990's (Ureña-Carrion et al. 2022).

Collections of sent and received letters are therefore stored in various archives for future generations to study. To enable Digital Humanities (DH) research (McCarty 2005; Gardiner and Musto 2015) on heterogeneous, distributed letter collections, data about the letters have been aggregated, harmonized, and provided for the research community through various databases and web services. Examples of such services include Europeana[1], Kalliope[2], The Catalogus Epistularum Neerlandicarum[3], Electronic Enlightenment[4], ePistolarium[5], the Mapping the Republic of Letters project[6], SKILLNET[7], correspSearch[8], and the Early Modern Letters Online (EMLO) catalogue[9].

From a technical point of view, epistolary metadata are challenging as letters are distributed in different cultural heritage organizations, have been catalogued using different data models and vocabularies, the letters are written in different languages, and the collections are typically incomplete. Using linked data provides a promising approach to tackle these problems. In (Tuominen et al. 2022) application of the idea to the Early Modern Letter Online database of the Oxford University was discussed. The LetterSampo Framework for publishing and using epistolary linked data for DH research was introduced in (Hyvönen, Leskinen, and Tuominen 2023) and (Leskinen et al. 2024), and later employed in the *Constellations of Correspondence - Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland* (CoCo) project[10] for developing LetterSampo Finland (1809–1917). This paper extends substantially our earlier paper on the CoCo project (Tuominen et al. 2022). We present results of applying and extending the LetterSampo Framework, using the so-called Sampo model (Hyvönen 2022) and Sampo-UI framework (Ikkala et al. 2022; Rantala et al. 2023), and the software developed for the case study regarding letters sent and received in Finland during the Grand Duchy in Finland (1809–1917).

2. Contributions of LetterSampo Finland (1809–1917)

There were many reasons for developing the LetterSampo Finland (1809–1917) system.

1. **Searching letters**. If a researcher was looking for the letters written or received by a person X, it has not been easy to find out in what archives such letter can be found. For

[1] http://www.europeana.eu
[2] http://kalliope.staatsbibliothek-berlin.de
[3] http://picarta.pica.nl/DB=3.23/
[4] http://www.e-enlightenment.com
[5] http://ckcc.huygens.knaw.nl/epistolarium/
[6] http://republicofletters.stanford.edu
[7] https://skillnet.nl
[8] https://correspsearch.net
[9] http://emlo.bodleian.ox.ac.uk
[10] CoCo project homepage in Aalto University: https://seco.cs.aalto.fi/projects/coco/

the first time, fundamental queries like "Find all letters sent (or received) by person X" can be answered under a single search engine in Finland.

2. **Providing a global view of correspondences**. The aggregated collection of LetterSampo Finland (1809–1917) provides a global view of the scattered letters can now be provided. From a quantitative point of view it has not been possible to answer simple questions, such as "How many letters available in archives are were sent in Finland during a period X in 1809-1917".

3. **Analyzing metadata**. Based on the metadata, it is now possible to analyze correspondences in flexible ways based on, e.g., persons, times, and places. For example: "How many letters did X receive from Y during time T", "How many letters were sent to place X by person Y?", "Who are the most active letter writers?".

4. **Analyzing letter content**. In some well-curated collection of prominent Finns, such as Johan Vilhelm Snellman (1806–1881), Zacharias Topelius (1818–1898), Johan Ludvig Runeberg (1804–1877), and Elias Lönnrot (1802–1884), not only metadata but also the letter contents are available in digital form for for textual and other analyses.

5. **Analyzing underlying networks**. Sending and receiving letters indicate social networks underlying the correspondences. Network analysis can be used to find out, for example, egocentric networks of individual people, social networks of groups of people, their central figures (hubs), and to study processional and family communications, or how the networks evolve in time (temporal networks).

6. **Developing infrastructure for epistolary data**. The primary data and metadata in archives is available in various formats, such as PDF and Word documents, spreadsheets, and in different kind of databases. It would be important to develop shared data models and vocabularies for representing epistolary data in the future based on the FAIR principles[11], so that the data would be more Findable, Accessible, Interoperable, and Re-usable in the future as the collections evolve and new ones are established.

7. **Analysing archival collections** The data can also be used for finding out what kind of epistolary fonds different archives have, how the collections have evolved in time, and to study geographical distributions of where the letters have been sent and received.

3. Paper outline

The full paper based on this abstract is organized as follows:

First, its is explained how a questionnaire was sent in Finland to over 100 CH organizations that were expected to host collections of letters from the time period c. 1800–1917. As a result, data from several organizations were received and statistics based on this investigation are provided in the paper.

[11]FAIR principles: https://www.go-fair.org/

Second, the tedious data cleaning process (Drobac et al. 2023) and pipelines for transforming primary dataset into linked open data are described. Several challenges were encountered: the data came in various heterogeneous forms that often needed human interpretation. Also issues of data quality, errors, and incomplete data arose. A major challenge here was linking and aligning person names with unique entities as person names change in time due to, e.g., marriages and deliberate name changes (Drobac, Leskinen, and Wahjoe 2023). Furthermore, various name variants have been used for the same persons in different archives and by different catalogers in different times. To tackle this, biographical data including, e.g., the times of living as well as the known name variations of individuals has been assembled from various data sources including earlier publications in Sampo series. Furthermore, the person data is also enriched from these external sources.

Third, the ontology-based data model extended from that of (Leskinen et al. 2024) and well as the vocabularies used for populating Linked Data are overviewed. In the data model the classes in the most central role are the metadata records, the letter resources, and the actors in correspondences.

From a data perspective, a major challenge in the case study was that in many, if not most cases, letter-wise metadata were not available but only metadata about archival units. For example, a particular unit in an archive may contain N letters that two families exchanged during a time period T, but it is not known who sent what letter to whom. On the other hand, in some cases pertaining to people of national importance, very detailed metadata about individual letters, including content annotated with mark-up such as TEI[12] was available. Another challenge of the data is its size: the KG contains information about nearly one million letters coming from 11 archives and 1100 fonds, with 95 000 historical people and 2000 places referred to.

Fourth, the process of establishing the LOD service and SPARQL endpoint using the Linked Data Finland platform LDF.fi[13] and publishing the data dumps an part of the national FIN-CLARIAH research infrastructure [14] are explained.

The LOD service SPARQL API can be used directly for DH research by, e.g., the Yasgui SPARQL query editor (Rietveld and Hoekstra 2017) or Jupyter Notebooks[15]. Examples with visualizations of this use case are given. Furthermore, we present results of using network analysis on epistolary data, using, e.g., the egocentric network based on Elias Lönnrot's correspondences (Poikkimäki, Leskinen, and Hyvönen 2024).

Fifth, the LetterSampo Finland (1809–1917) semantic portal with its four application perspectives based on integrating faceted semantic search and browsing with data-analytic tool is presented with examples for searching, exploring, and analyzing the underlying LetterSampo Finland (1809–1917) knowledge graph.

Finally, contributions of the research are summarized, challenges of using data-analytic methods in analyzing incomplete epistolary data are discussed, and directions for further research are proposed.

[12]Text Encoding Initiative TEI: https://www.tei-c.org/
[13]Linked Data Finland platform: https://ldf.fi
[14]Linked datata part of FIN-CLARIAH/DARIAH-FI: https://seco.cs.aalto.fi/projects/fin-clariah/
[15]Jupyter Notebooks: https://jupyter.org/

Acknowledgments

Thanks to Johanna Enqvist and Matti La Mela for their earlier contributions to the CoCo project. This research was funded by the Research Council of Finland and is part of the FIN-CLARIAH initiative that has received funding from the European Union NextGenerationEU instrument. Computing resources provided by the CSC – IT Center for Science were used in our work.

References

Drobac, Senka, Johanna Enqvist, Petri Leskinen, Muhammad Faiz Wahjoe, Heikki Rantala, Mikko Koho, Ilona Pikkanen, et al. 2023. "The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata." In *Digital Humanities in the Nordic and Baltic Countries Publication, DHNB2023 Conference Proceeding,* 5:248–262. 1. University of Oslo Library, Norway. https://doi.org/10.5617/dhnbpub.10669.

Drobac, Senka, Petri Leskinen, and Muhammad Faiz Wahjoe. 2023. "Navigating the Challenges of Deduplicating Actors in Historical Letter Exchanges." In *Proceedings of the 24th European Conference on Knowledge Management,* 24:1694–1697. 2. Academic Conferences International Limited. https://doi.org/10.34190/eckm.24.2.1317.

Gardiner, Eileen, and Ronald G. Musto. 2015. *The Digital Humanities: A Primer for Students and Scholars.* Https://doi.org/10.1017/CBO9781139003865. New York, NY, USA: Cambridge University Press.

Hotson, Howard, and Thomas Wallnig, eds. 2019. *Reassembling the Republic of Letters in the Digital Age.* Göttingen University Press. https://doi.org/10.17875/gup2019-1146.

Hyvönen, Eero. 2022. "Digital Humanities on the Semantic Web: Sampo Model and Portal Series." Accepted, *Semantic Web,* http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series.

Hyvönen, Eero, Petri Leskinen, and Jouni Tuominen. 2023. "LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters." *Journal on Computing and Cultural Heritage* 16 (1).

Ikkala, Esko, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2022. "Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces." *Semantic Web* 13 (1): 69–84. https://doi.org/10.3233/SW-210428.

Leskinen, Petri, Javier Ureña-Carrion, Jouni Tuominen, Mikko Kivelä, and Eero Hyvönen. 2024. "Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web." Under open review, *Semantic Web,* https://www.semantic-web-journal.net/content/knowledge-graphs-and-data-services-studying-historical-epistolary-data-network-science-1.

McCarty, Willard. 2005. *Humanities Computing.* Palgrave, London.

Miert, Dirk van. 2016. "What was the Republic of Letters? A brief introduction to a long history (1417–2008)." *Groniek* 204/205:269–287.

Poikkimäki, Henna, Petri Leskinen, and Eero Hyvönen. 2024. "Using Network Analysis for Studying Cultural Heritage Knowledge Graphs – Case Correspondence Networks in Grand Duchy of Finland 1809–1917." Under review. August. https://seco.cs.aalto.fi/publications/2024/poikkimaki-et-al-coco-2024.pdf.

Rantala, Heikki, Annastiina Ahola, Esko Ikkala, and Eero Hyvönen. 2023. "How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework." In *VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023*. CEUR Workshop Proceedings, Vol. 3508, October. https://ceur-ws.org/Vol-3508/paper3.pdf.

Rietveld, Laurens, and Rinke Hoekstra. 2017. "The YASGUI family of SPARQL clients." *Semantic Web – Interoperability, Usability, Applicability* 8 (3): 373–383. https://doi.org/10.3233/SW-150197.

Tuominen, Jouni, Mikko Koho, Ilona Pikkanen, Senka Drobac, Johanna Enqvist, Eero Hyvönen, Matti La Mela, Petri Leskinen, Hanna-Leena Paloposki, and Heikki Rantala. 2022. "Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland." In *6th Digital Humanities in Nordic and Baltic Countries Conference*. Forth-coming, preprint: https://seco.cs.aalto.fi/publications/2022/tuominen-et-al-coco-dhnb-2022.pdf. CEUR Workshop Proceedings, March.

Ureña-Carrion, Javier, Petri Leskinen, Jouni Tuominen, Charles van den Heuvel, Eero Hyvönen, and Mikko Kivelä. 2022. "Communications Now and Then: Analyzing the Republic of Letters as a Communication Network." In press, *Applied Network Science,* https://arxiv.org/abs/2112.04336v1.