LetterSampo Finland (1809–1917) Data Service and Portal: Searching, Exploring, and Analyzing Historical Letters and Their Underlying Networks

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

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
[3] Helsinki Insitute for Social Sciences and Humanities (HSSH), University of Helsinki
[4] Finnish Literature Society (SKS), Fnland

Abstract. Epistolary data are 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, 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 (1809–1917)* consisting of a Linked Open Data (LOD) service and a semantic portal targeted for Digital Humanities research.

Keywords: digital humanities · epistolary data · portals · data 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. This opportunity resulted into the so-called *Republic of Letters* that formed a basis for scientific thinking, values, and institutions in Early Modern times 1400–1800 [2].

Collections of sent and received letters are stored in various archives for future generations to study. To enable Digital Humanities (DH) research 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[5],

[5] `http://www.europeana.eu`

Kalliope[6], The Catalogus Epistularum Neerlandicarum[7], Electronic Enlightenment[8], ePistolarium[9], the Mapping the Republic of Letters project[10], SKILLNET[11], correspSearch[12], and the Early Modern Letters Online (EMLO) catalogue[13]. Epistolary metadata are challenging from a technical perspective 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.

Linked data provides a promising approach to tackle these problems. In [8] application of the idea to the EMLO database of the Oxford University was discussed. The LetterSampo Framework for publishing and using epistolary linked data for DH research was introduced in [3] and later employed in the CoCo project 2022-2025[14] [8] for developing the LetterSampo Finland (1809–1917) system, the topic of this paper. Our work seeks answers to the following research questions: RQ-1) How many and what kind of letters are there in Finnish fonds and collections from the Grand Duchy of Finland 1809-1917? RQ-2) How to make distributed heterogenous letter data FAIR for Digital Humanities research and applications? RQ-3) What kind of new historical insights can be obtained using DH methods on epistolary data and how?

2 LetterSampo Finland Data and LOD Service

To address RQ-1 a questionnaire was sent to over 100 Finnish archives for getting their data. For publishing and using the acquired data, an ontology-based data model was extended from our international LetterSampo system [3,6,9]. In the data model, the classes in the most central roles are the metadata records, the letter resources, and the actors in correspondences. The tedious data cleaning process and pipelines for transforming primary datasets into LOD are described in [1]. 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. To tackle the problem, 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

[6] http://kalliope.staatsbibliothek-berlin.de

[7] http://picarta.pica.nl/DB=3.23/

[8] http://www.e-enlightenment.com

[9] http://ckcc.huygens.knaw.nl/epistolarium/

[10] http://republicofletters.stanford.edu

[11] https://skillnet.nl

[12] https://correspsearch.net

[13] http://emlo.bodleian.ox.ac.uk

[14] CoCo project homepage in Aalto University: https://seco.cs.aalto.fi/projects/coco/

CH LOD publications in the Sampo series[15] systems. The person data was also enriched from these external sources.

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[16] was available. Another challenge of the data is its massive size: the KG contains information about over 1.2 million letters and their related entities.

Based on the harmonized data, a the LOD service and SPARQL endpoint using the Linked Data Finland platform LDF.fi[17] [4] was established as part of the national FIN-CLARIAH research infrastructure [18]. The LOD service SPARQL API can be used directly for DH research by, e.g., the Yasgui SPARQL query editor or Jupyter Notebooks. For example, results of using network analysis on epistolary data, using the egocentric network based on the correspondences of the polymath Elias Lönnrot are presented in [7].

3 LetterSampo Finland Portal

A portal that can be used without programming skills was then build on top of the LOD service. Based on the Sampo model and the Sampo-UI framework [5] for UI design, the landing page of the portal provides access to application perspectives where the instances of KG classes Letters (1 277 562 instances), Persons (116 462 instances), Fonds (1673 instances), and Places (2077 instances) can be searched using semantic faceted search where the facets correspond to the properties of the class. After filtering results by making selections on the facets, the result set can displayed as a table or using a variety of data-analytic tools and visualizations, such as charts, maps, and timelines. By selecting an instance from the result set, aggregated linked data related to it can be displayed and data-analyses and visualizations pertaining to the entity instance can be shown.

The data includes letters from four digital editions of prominent Finns, i.e., J. L. Snellman, E. Lönnrot, Z. Topelius, and A. Edelfelt, that include also the letter contents with man-made annotations for, e.g., topics discussed in the letter and places and people mentioned in them. For these cases specific application perspectives were provided supported by more advanced data for data analyses and visualization. For example, it is possible view letters on a map based on places mentioned or on charts visualizing the topics.

[15] Sampo series of over 20 CH LOD services and CH portals: https://seco.cs.aalto.fi/applications/sampo/

[16] Text Encoding Initiative TEI: https://www.tei-c.org/

[17] Linked Data Finland platform: https://ldf.fi

[18] Linked data part of FIN-CLARIAH/DARIAH-FI: https://seco.cs.aalto.fi/projects/fin-clariah/

4 Conclusions

LetterSampo Finland (1809–1917) system has provided novel answers to the research question RQ-1–3 set for the CoCo project. The portal is based on semantic web technologies, facilitate advanced Digital Humanities research, and the underlying LOD KG is massive in size.

References

1. Drobac, S., Enqvist, J., Leskinen, P., Wahjoe, M.F., Rantala, H., Koho, M., Pikkanen, I., Jauhiainen, I., Tuominen, J., Paloposki, H.L., Mela, M.L., Hyvönen, E.: The laborious cleaning: Acquiring and transforming 19th-century epistolary metadata. In: Digital Humanities in the Nordic and Baltic Countries Publication, DHNB2023 Conference Proceeding. vol. 5, pp. 248–262. University of Oslo Library, Norway (2023), https://doi.org/10.5617/dhnbpub.10669
2. Hotson, H., Wallnig, T. (eds.): Reassembling the Republic of Letters in the Digital Age. Göttingen University Press (2019), https://doi.org/10.17875/gup2019-1146
3. Hyvönen, E., Leskinen, P., Tuominen, J.: 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) (2023)
4. Hyvönen, E., Tuominen, J.: 8-star linked open data model: Extending the 5-star model for better reuse, quality, and trust of data. In: Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024). vol. 3759. CEUR Workshop Proceedings (September 2024), https://ceur-ws.org/Vol-3759/paper4.pdf
5. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web **13**(1), 69–84 (2022)
6. Leskinen, P., Ureña-Carrion, J., Tuominen, J., Kivelä, M., Hyvönen, E.: Knowledge graphs and data services for studying historical epistolary data in network science on the semantic web. Semantic Web (2024), https://www.semantic-web-journal.net/, under open review
7. Poikkimäki, H., Leskinen, P., Hyvönen, E.: Using network analysis for studying cultural heritage knowledge graphs – case correspondence networks in Grand Duchy of Finland 1809–1917 (August 2024), https://seco.cs.aalto.fi/publications/2024/poikkimaki-et-al-coco-2024.pdf, under review
8. Tuominen, J., Koho, M., Pikkanen, I., Drobac, S., Enqvist, J., Hyvönen, E., Mela, M.L., Leskinen, P., Paloposki, H.L., Rantala, H.: 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: DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference. pp. 415–423. CEUR Workshop Proceedings, Vol. 3232 (March 2022), http://ceur-ws.org/Vol-3232/paper41.pdf
9. Ureña-Carrion, J., Leskinen, P., Tuominen, J., van den Heuvel, C., Hyvönen, E., Kivelä, M.: Communications now and then: Analyzing the Republic of Letters as a communication network. Applied Network Science **7**(1) (2022). https://-doi.org/10.1007/s41109-022-00463-1