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 Institute for Social Sciences and Humanities (HSSH), University of Helsinki
[4] Finnish Literature Society (SKS), Finland

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 need to be aggregated, harmonized, and published as a global data service with APIs for Digital Humanities research and application development. This paper presents a new system *LetterSampo Finland – Historical Letters on the Semantic Web (1809–1917)* for publishing and studying epistolary data on a global level. The LOD service contains data about 1.2 million letters sent or received in the Grand Duchy of Finland during 1809–1917, aggregated from 13 Finnish CH organizations, harmonized by using a shared ontological data model and vocabularies. A new semantic portal has been created on top of the LOD service. This portal can be used to search, explore, and analyze letters, letter collections, and networks of correspondents in 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 in the so-called *Republic of Letters* that formed a basis for scientific thinking, values, and institutions in Early Modern times 1400–1800 [15,5].

Collections of sent and received letters are stored in various archives for future generations to study. To enable Digital Humanities (DH) research [4] 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], 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 cataloged 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 [21] an application of the idea was presented to the EMLO database of the University of Oxford. The LetterSampo Framework for publishing and using epistolary linked data for DH research was introduced by the international system *LetterSampo – Historical Letters on the Semantic Web*[14] in [9,14], applied using network analysis [22,14], and later employed in the CoCo project 2021–2025[15] [20] for the LetterSampo Finland system. 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 heterogeneous letter data Findable, Accessible, Interoperable, and Re-usable (FAIR) for DH research and applications? RQ-3) What kind of new historical insights can be obtained using DH methods on epistolary data and how? Our research hypothesis is to re-use, adapt, and extend the LetterSampo Framework to create LetterSampo Finland: a new Knowledge Graph (KG), data service, and a semantic portal on top of it.

The KG introduced in this paper is openly available CC BY 4.0 on the Linked Data Finland platform[16] with a SPARQL endpoint, and as a data dump on Zenodo.org[17]. The portal is available online at `https://kirjesampo.fi`.[18]

[5] `http://www.europeana.eu`
[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] Portal available at: `https://lettersampo.demo.seco.cs.aalto.fi/en/`
[15] CoCo project technical homepage: `https://seco.cs.aalto.fi/projects/coco/`
[16] LetterSampo Finland LOD service: `https://ldf.fi/dataset/coco/`
[17] Data dumb in RDF: `https://doi.org/10.5281/zenodo.15210589`
[18] The LOD and portal are available after the publication event on May 27, 2025: `https://seco.cs.aalto.fi/events/2025/2025-05-27-kirjesampo/`

2 LetterSampo Finland Data and LOD Service

To address the research question 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 [9,14,22]. In the data model, the classes in the most central roles are the metadata records of the letters and the actors in correspondences. The tedious data cleaning process and pipelines for transforming primary datasets into LOD are described in [2,1]. Several challenges were encountered: the data came in various heterogeneous forms that often needed human interpretation. In addition, issues arose concerning data quality, errors, and incomplete data. A major challenge here was linking and aligning person names with unique entities as person names change in time due, e.g., marriages and deliberate name changes. To tackle the problems, biographical data including, e.g., the times of living, as well as the known name variations of people, have been assembled from various data sources including earlier CH LOD publications in the Sampo series[19] 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. However, in some cases, pertaining to people of national importance, very detailed metadata about individual letters was available, including markup-annotated content such as TEI[20]. Another challenge of the data is its massive size: the KG contains information about ca. 1.28 million letters and related entities.

Based on the harmonized data, an LOD service and a SPARQL endpoint using the Linked Data Finland platform LDF.fi[21] [6,10] was established as part of the national FIN-CLARIAH research infrastructure[22] [8]. The LOD service SPARQL API can be used directly for DH research by, e.g., the Yasgui SPARQL query editor[23] [19] or Jupyter Notebooks, and for developing applications, such as portals.

3 LetterSampo Finland Portal

A portal that can be used without programming skills has been built on top of the LOD service. Based on the Sampo model [7] and the Sampo-UI framework

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

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

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

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

[23] `https://yasgui.triply.cc/`

[11,18] for UI design, the landing page of the portal provides access to application perspectives where the instances of the KG classes Letters (1 281 936 instances), Persons and organizations (118 076 instances), Fonds (1689 instances), and Places (2165 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. V. Snellman (1806–1881), E. Lönnrot (1802–1884), Z. Topelius (1818–1898), and A. Edelfelt (1854–1905). These data contain letter texts with possible man-made annotations for, e.g., keywords for topics discussed in the letters and mentioned places and people in the texts. For these editions, four specific application perspectives were implemented and supported by more advanced search features and visualizations using the annotations. For example, one can view letters on a map based on places mentioned in them or on charts visualizing the topics.

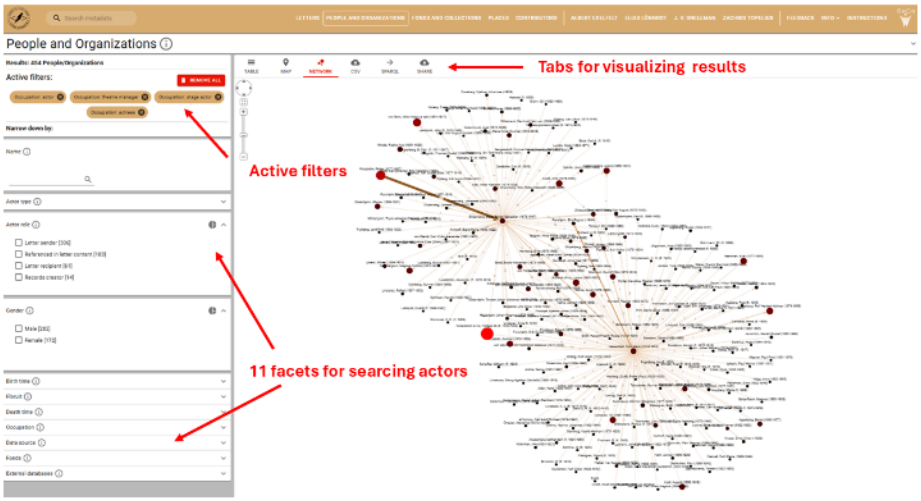

Fig. 1: Correspondence network of people related to theater world

The landing page also provides direct example links to pages based on faceted search and visualizations. For example, “Letters between Johan Ludvig and Fredrika Runeberg”, “Timeline of Helene Schjerfbeck’s letters”, “A. Edelfelt’s letters on a map”, and “Correspondence network of people related to theater world“ depicted in Fig. 1. Here the end user has selected the actor perspective

PEOPLE AND ORGANIZATIONS. On the left several facets are shown to filter out actors by Name, Actor type, Actor role, Gender, Birth time, Floruit, Death time, Occupation, Data source, Fonds, and External databases. Active filters to find theater people based on a selection of occupations can be seen above the facets. The search results are presented on the right using different visualization tabs: TABLE (list of results), MAP (letters on a map), and NETWORK that visualizes the correspondence networks of the filtered actors. On the tab CSV it is also possible to see the results in CSV format and the tab SPARQL leads to a Yasgui page where the filtering SPARQL search query can be edited for further studies.

4 Contributions and Discussion

The LetterSampo Finland system has provided novel answers to the research questions RQ-1–3 set in Section 1:

RQ-1. How many and what kind of letters are there in Finnish fonds and collections from the Grand Duchy of Finland 1809-1917? Although it was not possible to get letter data from all potential archives in Finland, it seems that the 1.28 million letters harvested cover most of the Finnish letter data (1809–1917) available in professional archives, which is more than was initially expected.

RQ-2. How to make distributed heterogeneous letter data FAIR for DH research and applications? The data harvesting and cleaning pipeline [2] was tedious, required manual work, and the metadata available was in many cases incomplete and uncertain. However, the KG created is useful, if the properties and limitations of the datasets are made transparent for the end users to support data literacy [12]. The LetterSampo Framework including the Sampo model and Sampo-UI used for the implementation worked well and were deemed re-usable.

RQ-3. What kind of new historical insights can be obtained using DH methods on epistolary data and how? First studies have been conducted to provide new insights for research in humanities: The possibilities of using LetterSampo Finland in research are discussed in [3] and [16]. Network analyses using, e.g., the egocentric network based on the correspondences of the polymath Elias Lönnrot are reported in [17]. Article [13] explores how critical data modeling and the application of data science methods can be used to mitigate archival biases when working with big historical data, in this case the LetterSampo Finland letter collection KG.

Further research on LetterSampo Finland includes more formal evaluation of the system from different perspectives, such as data quality, fitness for research use, and usability of the user interface of the portal.

Acknowledgments 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 services provided by the CSC – IT Center for Science were used.

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. Digital Humanities in the Nordic and Baltic Countries Publications **5**(1), 248–262 (2023). https://doi.org/10.5617/dhnbpub.10669
2. Drobac, S., Leskinen, P., Wahjoe, M.F.: Navigating the challenges of deduplicating actors in historical letter exchanges. European Conference on Knowledge Management **24**(2), 1694–1697 (2023). https://doi.org/10.34190/eckm.24.2.1317
3. Enqvist, J., Pikkanen, I.: Kirjeluettelot tutkimusaineistona ja kulttuuriperintönä: metadatan mahdollisuudet digitaalisen käänteen jälkeen. In: Karhu, H., Kivilaakso, K., Parente-Čapková, V. (eds.) Tutkimuspolkuja yksityisarkistoihin – Aineistot historian, kulttuurin ja kirjallisuuden tutkimuksessa. pp. 390–426. Suomalaisen Kirjallisuuden Seura, Helsinki (2024)
4. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015), `https://doi.org/10.1017/CBO9781139003865`
5. 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`
6. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer-Verlag (2014), `https://doi.org/10.1007/978-3-319-11955-7_24`
7. Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web **14**(4), 729–744 (2022). https://doi.org/10.3233/SW-223034
8. Hyvönen, E.: How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web. Semantic Web **15**(4), 1499–1513 (2024). https://doi.org/10.3233/SW-243468
9. 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). https://doi.org/10.1145/3569372
10. 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`
11. 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). https://doi.org/10.3233/SW-210428
12. Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science **49**(1), 3–14 (2015). https://doi.org/10.1177/0961000615616450
13. La Mela, M., Pikkanen, I., Paloposki, H.L., Tuominen, J.: A critical collection history of nineteenth-century women's letters. Overcoming the occluded archive with data-driven methods. Digital Humanities Quarterly (Forth-coming, 2025), revised manuscript for DHQ Special Issue 'Data Science and History: Practicing and Theorizing Data-Driven Inquiries into the Past'

14. 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
15. van Miert, D.: What was the Republic of Letters? A brief introduction to a long history (1417–2008). Groniek **204/205**, 269–287 (2016)
16. Paloposki, H.L., Pikkanen, I.: Learning to read digital? constellations of correspondence project and humanist perspectives on the aggregated 19th-century Finnish letter metadata. extended abstract. In: Baudry, J., Burkart, L., Joyeux-Prunel, B., Kurmann, E., Mähr, M., Natale, E., Sibille, C., Twente, M. (eds.) Digital History Switzerland 2024: Book of Abstracts (2024), https://digihistch24.github.io/submissions/444/
17. Poikkimäki, H., Leskinen, P., Hyvönen, E.: Exploring Cultural Heritage Knowledge Graphs – Case of Correspondence Networks in Grand Duchy of Finland 1809–1917. Digital Humanities in the Nordic and Baltic Countries Publications **7**(2) (2025). https://doi.org/10.5617/dhnbpub.12289
18. Rantala, H., Ahola, A., Ikkala, E., Hyvönen, E.: 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. vol. 3508. CEUR Workshop Proceedinbgs (2023), https://ceur-ws.org/Vol-3508/paper3.pdf
19. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability **8**(3), 373–383 (2017). https://-doi.org/10.3233/SW-150197
20. 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
21. Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., Hotson, H.: Reassembling the Republic of Letters - a linked data approach. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 76–88. CEUR Workshop Proceedings, vol. 2084 (March 2018), http://www.ceur-ws.org/Vol-2084/paper6.pdf
22. 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