LetterSampo Finland Knowledge Graph, Data Service, and Semantic Portal for Researching Epistolary Data of the Grand Duchy of Finland (1809–1917)

Eero HYVÖNEN [a,b], Petri LESKINEN [a], Henna POIKKIMÄKI [a], Heikki RANTALA [a], Annastiina AHOLA [a], Rafael LEAL [a], Jouni TUOMINEN [b,c], Senka DROBAC [b], Ossi KOHO [b], Ilona PIKKANEN [d], Hanna-Leena PALOPOSKI [d]

[a] *Aalto University, Department of Computer Science*
[b] *University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)*
[c] *University of Helsinki, Helsinki Institute for Humanities and Social Sciences (HSSH)*
[d] *Finnish Literature Society*

Abstract. Epistolary letter data is by nature stored in geographically distributed archives and collections, as letters are sent between different people and places. To get a global view and analyze correspondences, archive collections, and underlying egocentric and social networks, data from separate data silos in different cultural heritage (CH) organizations must be aggregated, harmonized, and published as a global data service with APIs for Digital Humanities research and application development. This paper presents the system *LetterSampo Finland* consisting of a Linked Open Data (LOD) service and a semantic portal designed for these purposes. The LOD service contains metadata about nearly 1.3 million letters, 120 000 people and organizations in the Grand Duchy of Finland during 1809–1917, aggregated from sixteen Finnish CH organizations and 1700 fonds, 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 the Sampo-UI framework, 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 between the correspondents.

Keywords. linked data, epistolary data, 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 to send and receive letters across Europe and beyond, based on a revolution in postal services. This opportunity resulted into the so-called *Republic of Letters* (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 [1,2]. Sending letters in many ways analogous to modern means of communication using the Internet, email, social media, and the World Wide Web (WWW) since the 1990's [3].

Collections of letters have been stored in various archives for future generations to study. This paper presents a new approach, LOD service, and portal for publishing and using epistolary data for digital humanities research, based on semantic web technologies. As a case study, Finnish 19th century letters are considered.

Related works To enable Digital Humanities (DH) research [4,5] 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]. However, these online systems provide letter data mostly for humans to read but not as data needed for computational analyses of Digital Humanities. However, in some cases, such as the web service of correspSearch [6], an API is also provided, in this case for TEI-XML data. The idea of using Linked Data for publishing and using epistolary data was discussed in [7] in relation to EMLO. The Norwegian Correspondences project [8] aim to base their web services on linked data. Network analysis of based on epistolary data is discussed, e.g., in [9,10]

From a technical point of view, epistolary metadata are challenging 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 and contain uncertain data. Using linked data provides a promising approach to tackle these problems. In [11] application of using a linked data approach to the Early Modern Letter Online database of the Oxford University was discussed with some demonstrational data analyses. The LetterSampo Framework for publishing and using epistolary linked data for DH research was introduced in [12] and [13], with LOD published in a SPARQL endpoint and prototype LetterSampo online. This framework, extending the so-called Sampo model [14] and Sampo-UI framework [15,16], was later employed in the *Constellations of Correspondence - Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland* (CoCo) project[10] (2021–2025) for developing the LETTERSAMPO FINLAND system. This paper overviews and extends substantially our earlier papers on the CoCo project [11,17,18,19,20,21]. We present the final public LOD service and portal demonstrator of the project with first results of using them for researching epistolary data of the Grand Duchy in Finland (1809–1917). The Knowledge Graph (KG) introduced in this paper is openly available CC BY 4.0 on the Linked Data Finland platform[11] with a

[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/
[11] LetterSampo Finland LOD service: https://ldf.fi/dataset/coco/

SPARQL endpoint, and as a data dump on Zenodo.org[12]. The portal is available online at `https://kirjesampo.fi`.[13]

The paper is organized as follows. Motivations for creating the LetterSampo from a DH point of view are first listed in the next Section 2. After this process of creating the LOD service for data analyses and the portal on top of is explained (Section 3), and major semantics challenges are discussed (Section 4). In Section 5 it is the shown, how the LOD service can be used for research, and in Section 6 using the portal for this task is explained. Finally, contributions of the research are summarized, challenges of using data-analytic methods for analyzing incomplete epistolary data are discussed, and directions for further research are proposed (Section 7).

2. Motivating Use Cases and Research Questions

There were many reasons for developing the LetterSampo system.

1. **Searching letters**. If a researcher is looking for particular letters, say written or received by a person *x*, it is not easy to find out in what archives such letters can be found. For the first time, fundamental queries like "Find all letters sent or received by a person *x*" can be answered under a single search engine in Finland.
2. **Providing a global view of correspondences**. The aggregated collection of LetterSampo provides a global view of the locally distributed letter collections. From a quantitative point of view it has not been possible to answer simple questions, such as "How many letters are there in the archives that were sent in Finland during a period *t* in the 19th century".
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 *p* by person *y*?", "Who are the most active letter writers?".
4. **Analyzing letter content**. In some well-curated collections of prominent people not only metadata but also the letter contents are available in digital form for textual and other computational analyses. In our case, we got access to the digital letter editions of the Finns Johan Vilhelm Snellman (1806–1881), Zacharias Topelius (1818–1898), Albert Edelfelt (1854–1905), and Elias Lönnrot (1802–1884).
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 kinds of databases. It would be important to develop shared data models and vocabularies for representing epistolary data in the

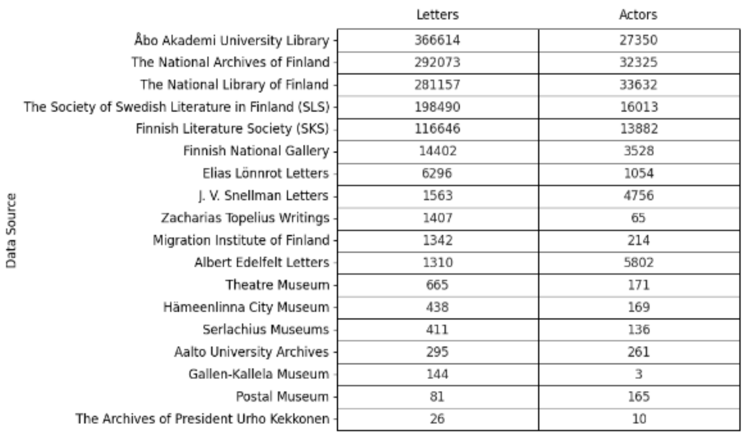
[12] Data dumb in RDF: `https://doi.org/10.5281/zenodo.15210589`

[13] 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/`

future based on the FAIR principles[14], 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 organizations have, how the collections have evolved in time, and to study geographical distributions of where the letters have been sent and received.

3. Creating the LetterSampo Data Service and Portal



Data Source	Letters	Actors
Åbo Akademi University Library	366614	27350
The National Archives of Finland	292073	32325
The National Library of Finland	281157	33632
The Society of Swedish Literature in Finland (SLS)	198490	16013
Finnish Literature Society (SKS)	116646	13882
Finnish National Gallery	14402	3528
Elias Lönnrot Letters	6296	1054
J. V. Snellman Letters	1563	4756
Zacharias Topelius Writings	1407	65
Migration Institute of Finland	1342	214
Albert Edelfelt Letters	1310	5802
Theatre Museum	665	171
Hämeenlinna City Museum	438	169
Serlachius Museums	411	136
Aalto University Archives	295	261
Gallen-Kallela Museum	144	3
Postal Museum	81	165
The Archives of President Urho Kekkonen	26	10

Figure 1. Data providers of LetterSampo Finland with the number of letters (some 1290000 in total) and related actors (people and organizations, some 118000 in total) in their collections

Acquiring data from participating organizations Our project started by sending a questionnaire to over 100 CH organizations in Finland that were expected to host collections of letters from the time period of the Grand Duchy of Finland 1809–1917. As a result, data from the organizations and collections listed in Fig. 1 were finally received. The massive amount of data exceeded the initial expectations.

The data include not only original letter data from archives, libraries and museums, but also already aggregated and enriched data form 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). This meant that the same letters could appear multiple times in the data and had to be deduplicated. The digital edition data include the textual letter contents 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 provided and supported by more advanced search features and visualization using the annotations. For example, one can view letters on a map based on places mentioned in them or on charts visualizing the topics.

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

Data model For publishing and using the data, an ontology-based data model depicted in Fig. 2 was designed by extending that of our earlier international LetterSampo system [12,13,3]. In the data model, the classes in the most central roles are the metadata records of the letters (:Letter), actors (people and organizations) (crm:E30_Actor), collections and fonds (:Fond), and places (crm:E53_place) related to the correspondences. They were later used as bases for the application perspectives in the LetterSampo Finland portal to presented in Section 6. For representing actors, the proxy data model of Europeana [22] was employed in order keep track of possibly conflicting data from different archival sources. More information and documentation of the data model can be found at the LOD service homepage on the Linked Data Finland platform.

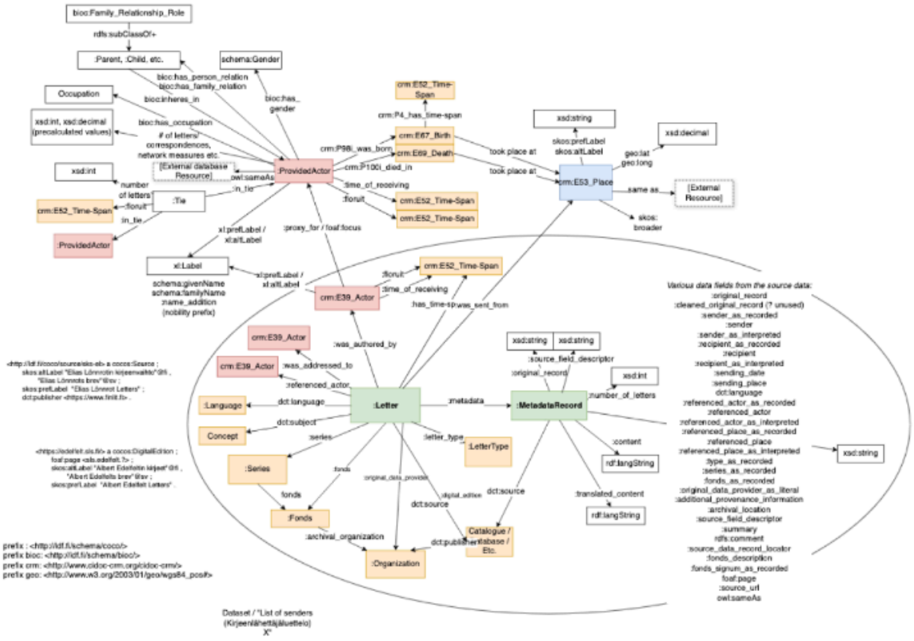


Figure 2. Data model for LetterSampo

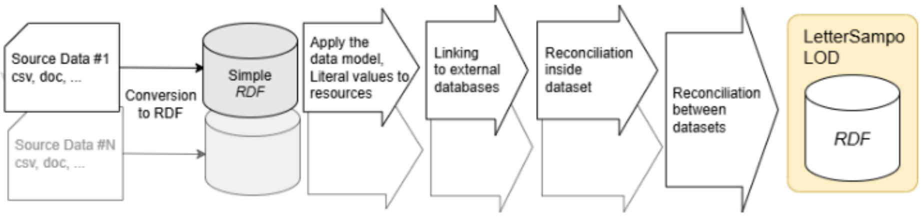

Source Data #1
csv, doc, ...
Source Data #N
csv, doc, ...
Conversion to RDF
Simple
RDF
Apply the data model, Literal values to resources
Linking to external databases
Reconciliation inside dataset
Reconciliation between datasets
LetterSampo LOD
RDF

Figure 3. Pipeline for transforming data of memory organizations into LOD

Data cleaning, transformation, and Linking pipeline The data cleaning and transforming pipeline from data providers to LetterSampo LOD service is depicted in

Figure 3. The tedious data cleaning process [17] involved first transforming primary datasets into simple RDF form that was then enriched by data linking of literal values to each other internally and externally to related linked datasets. The new SampoSampo data linking service [23], a VIAF.or-like [24] mapping service for Sampo systems and various external datasets was used here, and lots of links the following datasets were found (number of links in parentheses): Geneanet genealogy service[15] (18447), Wikidata.org (14071), Wikipedia.org (8966), AcademySampo [25,26] about Finnish academicians (1640–1899) (8378), BiographySampo [27,28] of National biography of Finland (6139), Albert Edelfelts brev letter edition (5827), Kanto[16] registry of the National Library (5567), Wikitree.org genealogy service (1843), BookSampo [29] on Finnish Fiction Literature (1179)], Union List of Artist Names ULAN of Getty Ressearch (1063), ParliamentSampo [30] of Parliament of Finland data (683), and OperaSampo [31] of historical music theater performances (160). For some 85 000 people, no external links were found, suggesting that most of the actors in the data are not well-known nationally.

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 [18]. Furthermore, various name variants have been used for the same persons in different archives and by different catalogers in different times.

To solve data linking 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[17] systems. AT the same time, the actor data was also enriched from these external sources.

LOD service online Finally, the data was published as a LOD service and SPARQL endpoint using the Linked Data Finland platform LDF.fi [32,33]. 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 [34] or Jupyter Notebooks[19]. The LOD service can also be used as a basis for developing applications such as semantic portals. It turned out that the Apache Jena Fuseki SPARQL server[20] in use by default in LDF.fi was not efficient enough for dealing the massive about of letter instances in LETTERSAMPO FINLAND. As a remedy, the new open source QLever SPARQL engine[21] turned out to be an order of magnitude faster and was used.

[15] Geneanet: `https://fi.geneanet.org/`

[16] Kanto: `https://finto.fi/finaf/en/`

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

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

[19] Jupyter Notebooks: `https://jupyter.org/`

[20] Fuseki SPARQL server: `https://jena.apache.org/documentation/fuseki2/`

[21] QLever SPARQL engine: `https://pypi.org/project/qlever/`

4. Semantic Challenges of Uncertain and Incomplete Data

In our earlier work on developing the international LetterSampo system [12,13], high quality data about individual letters from Oxford EMLO database, CKCC corpus from the Huygens Institute, and correspSearch data from Berlin Brandenburg Academy of Sciences was available, as the letters were between identified prominent people, such as Isaac Newton (1643–1727), Gottfried Leibniz (1646–1716), and others. In LETTERSAMPO the metadata is not always so accurate and included lots of letters between less well-known actors, which set new semantic challenges for our work.

From a data perspective, a major challenge in LETTERSAMPO was that in many cases letter-wise metadata was not available but only metadata about correspondences in more general level. For example, a particular unit in a fonds may contain a set of letters that two families exchanged during a certain time period, but it is not known who sent what letter to whom. Detailed metadata on individual letters was typically available only in cases pertaining to people of great national importance or from some participating organizations. Another challenge of the data was their size: automatic methods had to be used without the possibility of manual corrections.

A limitation in many epistolary systems, such as correspSearch, is that only metadata about letters is available, not their textual content or metadata about it. The same limitation applies to most letters is LETTERSAMPO, too. However, the data includes four digital editions in which the textual content is available, and is some cases also metadata about it, possibly with mark-up, such as TEI[22].

Semantic challenges were encountered concerning modeling the correspondents, when the data is used for searching and analyzing letters. When using faceted search, several categories are needed for categorizing the sender/receiver when filtering results. For example, the end user may want to filter out only letters whose sender and receiver are known persons with a known gender. Furthermore, the sender and receiver are not necessarily people but also:

1. *Groups of people*, such as married couples or families. For example, assume the metadata may only tells that a number of letters was exchanged between two families, but letter-wise metadata is not available. Then the letters cannot be searched and analyzed by the gender of the sender/receiver, as both female and male correspondents may be included in a family.
2. *Organizations*, such as companies, public offices, or newspapers. Unlike people, organizations do not have gender, place of birth/death, and time of birth/death, and cannot be searched or analyzed along these dimensions.

Uncertainty regarding the sender/receiver or other entity has to be represented in some way. The following categories of uncertainty could be identified and are used in the portal for faceted search:

1. *Missing value*. Metadata value is missing in the metadata in the one dataset although it may be there in other datasets. For example, in some datasets the places to where letters are sent or the type of the letter are not included in the metadata.
2. *Unknown value*. Sometimes a metadata element is expected to have a value, but it is not known. For example, assume that the sender is marked in the primary

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

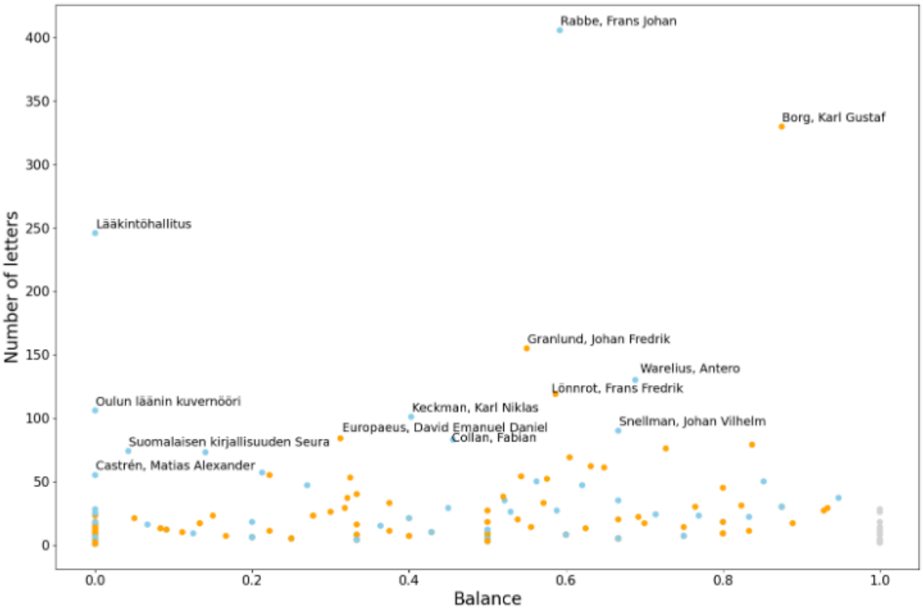


Figure 4. Number of letters and balance between Lönnrot and his correspondents

data by name but the gender is not given. The gender is then implicitly given and not missing, but cannot necessarily be determined using the name and remains unknown.

3. *Unidentified value*. In this case, there is a value in the metadata, but its value cannot be understood by the machine. For example, the value may be formatted in a way that cannot be identified or contains a confusing typo.
4. *Not applicable value*. The value is not related to a sender/receiver or other entity. For example, organizations may send/receive letters like people, but do not have a gender. The value is then considered not Missing, Unknown, or Unidentified, but Not applicable.

5. Using the LOD Service

The LetterSampo Finland LOD service can be used for data analyses directly. The open SPARQL endpoint is then used for querying a subset of data of interest and after that it can be analyzed and visualized by using, e.g., Python scripting and libraries in Jupyter notebooks.

Examples of this idea are presented in [19]. For example, Fig. 4 shows on the y-axis the number of letters exchanged between Mr. Elias Lönnrot and other letter writers. The x-axis describes the balance between Lönnrot and other letter writers. The balance between two actors is calculated by dividing the minimum number of letters one actor has sent to another by the maximum number of letters that one actor has sent to another [35]. Balance of 0.0 tells that letters have been sent only in one direction and when balance is 1.0 actors have sent equally many letters to each other. In the figure, the blue node color tells that Lönnrot has sent all or most of the letters that have been sent between Lönnrot and the other letter writer, and orange tells that the other writer has sent most of the letters. The balance of 1.0 is represented by the gray color in the figure.

In the upper right corner there are actors who sent many letters to Lönnrot and also received many letters from Lönnrot, such as Frans Johan Rabbe, who was also a doctor and worked for "Lääkintöhallitus" (Health Institute of Finland) during that time, and Carl Gustaf Borg, who worked during the same time at the University of Helsinki as Lönnrot. In the upper left side, where the correspondence is unbalanced, there are groups like the "Suomalaisen Kirjallisuuden Seura" (Finnish Literature Society) for which Lönnrot has sent a lot of letters. A low balance value can also reveal potentially missing letters. For example, according to Fig. 4 Lönnrot has sent more than 50 letters to his friend, linguist and explorer Matias Alexander Castrén, but how likely is it that Castrén never answered his letters? Deeper analyses are needed with close reading, but the point here is that computationally obtained analyses can alert the humanist researcher about potentially interesting historical phenomena for further study.

AS an other use case of the LETTERSAMPO FINLAND LOD service, in [21] a "virtual archive" of women's epistolary exchange in 19th-century Finland is presented. Here it is studied by quantitative analysis, enriched metadata, and network visualizations using SPARQL and Jupyter notebooks, e.g., whether women are archival protagonists, or are their materials embedded within the collections of male relatives? Or do the data reveal overlooked women with extensive archival networks absent from historical narratives?

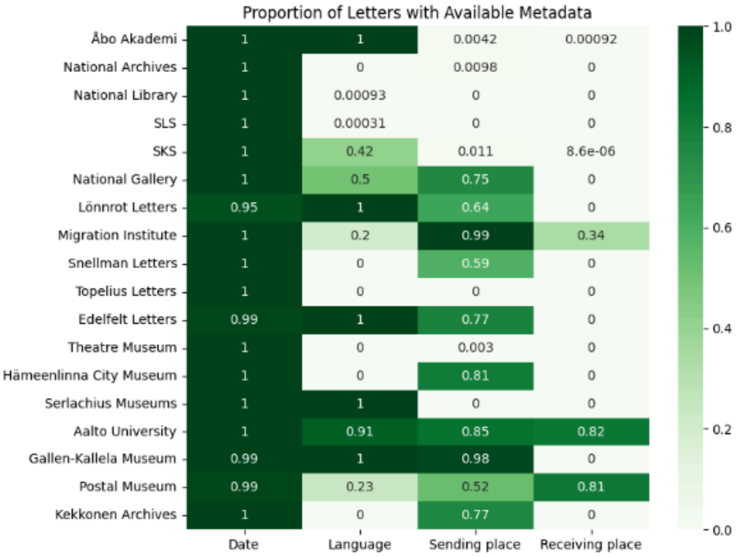


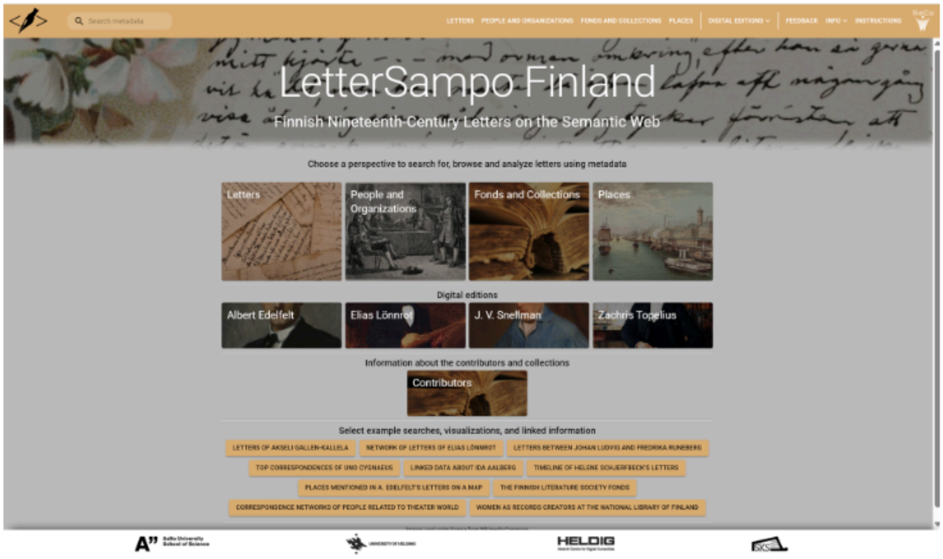
Figure 5. Analysis of metadata quality in the data sources of LETTERSAMPO FINLAND

The LOD service is also useful for analyzing the characteristics and quality of the LOD. For example, Fig. 5 shows the proportion of letters in each data source that have information about date, language, sending and receiving place. Almost all letters have some estimate about the time of sending, although the exact sending date is rarely known.

The availability of the language of the letter differs from source to source. The place of sending is better known in the smaller data sources, and the place of receiving the letter is usually not known. The sending place can vary from the sender's residence to country, but is usually a village, town, or city.

6. Using the Portal for Searching, Browsing, and Analyzing Data

After establishing the LOD service, a portal was built on top of the SPARQL endpoint. The portal can be used for epistolary research without programming skills by researchers and the general public. Based on the Sampo model [36] and the Sampo-UI framework [37,38] for UI design, the portal provides means for searching, browsing, analyzing, and visualizing instances of KG core classes.



LetterSampo Finland

Finnish Nineteenth-Century Letters on the Semantic Web

Choose a perspective to search for, browse and analyze letters using metadata

Digital editions

Information about the contributors and collections

Select example searches, visualizations, and linked information

Search metadata
LETTERS
PEOPLE AND ORGANIZATIONS
FONDS AND COLLECTIONS
PLACES
DIGITAL EDITIONS
FEEDBACK
INFO
INSTRUCTIONS
Letters
People and Organizations
Fonds and Collections
Places
Albert Edelfelt
Elias Lönnrot
J. V. Snellman
Zachris Topelius
Contributors
LETTERS OF AKSELI GALLEN-KALLELA
NETWORK OF LETTERS OF ELIAS LÖNNROT
LETTERS BETWEEN JOHAN LUDVIG AND FREDRIKA RUNEBERG
TOP CORRESPONDENCES OF UNO CYGNAEUS
LINKED DATA ABOUT IDA AALBERG
TIMELINE OF HELENE SCHJERFBECK'S LETTERS
PLACES MENTIONED IN A. EDELFELT'S LETTERS ON A MAP
THE FINNISH LITERATURE SOCIETY FONDS
CORRESPONDENCE NETWORKS OF PEOPLE RELATED TO THEATER WORLD
WOMEN AS RECORDS CREATORS AT THE NATIONAL LIBRARY OF FINLAND
HELDIG

Figure 6. Landing page of LetterSampo Finland

Main Functionalities of the Portal The landing page of a Sampo portal contains a series of *application perspective* windows that allow one to search, browse, and analyze the underlying KG from different perspectives, based on the classes of the KG. A perspective for a class contains a faceted search engine whose facets are based on the properties of the class; by making selections on the facets a corresponding subset of individuals of the class is retrieved as the search result, and hit counts on the facet categories are updated to guide the search. The result set can then be analyzed on different tabs, e.g., on a map or timeline or using graphs. At any point, an individual can be chosen for a closer look at its *instance page* that provides comprehensive linked data about the individual. Here it is also possible to analyze and visualize data about the individual by using separate tabs. In the case of LetterSampo Finland, the are four main application perspectives available for 1) Letters, 2) People and organizations, 3) Finds and collections, and 4) Places (upper row of boxes in Fig. 6). They can be used for searching all data in the KG.

The landing page also contains—under the main perspectives in Fig. 6—separate focused search views into the data based on four previously published digital editions: Albert Edelfelt, Elias Lönnrot, J. V. Snellman, and Zachris Topelius. They provide not only metadata but also the contents of the letters.

Below these, there is still one view that provides a listing and links to the 15 memory organizations that originally collected and own the LETTERSAMPO data.

At the bottom of the landing page there are a number of buttons that can be used to activate a number of example searches. For example: “Letters between Johan Ludvig Runeberg and Fredrika Runeberg” or “Correspondence networks of people related to theater world”.

The top bar of the landing page is visible in all views and provides shortcut buttons for directly switching to any application perspective. First, there is a global search window for the KG metadata. The FEEDBACK button can be used to send feedback to the system developers, for example, new development ideas and reports of possible errors. The INFO button provides additional information about the LETTERSAMPO FINLAND project and the research behind it, as well as more detailed data analyses of the portal’s data for digital humanities research, such as Fig. 5. These analyses aim to make the data more transparent and support data literacy of the end user when performing data analyses. Finally, the INSTRUCTIONS button links to instructions for use.

The anticipated main use case of the portal is to find letters and data related to them. In below, a brief introduction to using the portal’s search views and data analytics tools is presented.

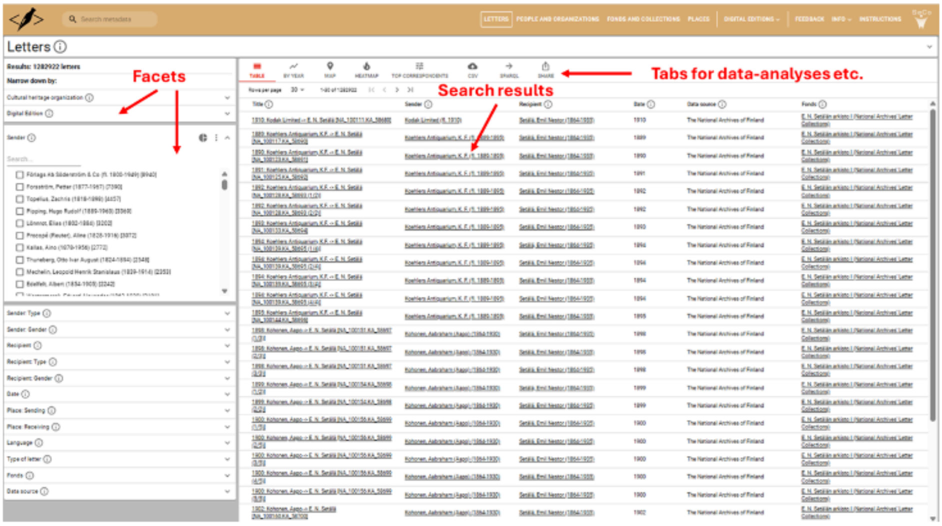


Figure 7. Letters perspective with faceted search and tabs for data-analyses

Letters Perspective The LETTERS perspective provides 14 facets for filtering letters, such as Sender, Receiver, Type of Letter, etc. Fig. 7 depicts this perspective with the search facets on the left, the Sender facet open, and the search result on the right.

The hit counts for the categories in the Sender facet show the number of letters sent by individuals and organizations in order of magnitude, with Förlags A Söderstöm & Co at the top with almost 9000 letters sent. The facet categories can alternatively be displayed alphabetically and can be searched using the facet's internal search function.

The search result on the right is presented in table format by default on the TABLE tab, but by changing the tab, the result can be visualized and analyzed, for example, on maps, or by identifying the most active correspondents in the search result.

The facets allow one to search and view the number of letters, for example, by the sender or receiver facet. Each facet lists the possible search options as a set of categories (e.g., different people). For example, when one selects a person from the sender facet, such as Elias Lönnrot, one will get letters sent by Lönnrot as the search result. Each category is associated with a numerical hit count, which indicates the number of search results (in this case, the number of letters) if that category is selected next. The hit counts of all facets are automatically updated after each search selection.

Hit counts help the user to direct the search and prevent making choices that would lead to a dead end, i.e., an empty search result. Hit counts also provide the opportunity to statistically analyze the distribution of search results in relation to the categories of different facets. The distribution can be visualized by clicking on the pie symbol at the beginning of a facet, which visualizes the distribution of the current result set as a pie chart in relation to that facet. Dynamically calculated hit numbers as well as the possibility of using hierarchical facets make facet search more versatile than traditional filtering of search results with search criteria.

After each facet selection, the result field on the right side of the screen is updated and presented on the following tabs:

1. The TABLE view displays the results in a list.
2. BY YEAR shows the annual number of letters as a line graph. The zoom tools in the top right corner allows one to view more detailed periods on a graph.
3. MAP visualizes the locations where letters are written or received.
4. HEATMAP visualizes the writing and receiving locations as a heat map (the redder the color, the more letters).
5. TOP CORRESPONDENTS displays the 20 largest correspondences on a time axis. Note that most correspondences only have a start and end year (e.g. 1860–1895). In such cases, the letters of the correspondence in question are distributed uniformly over the period in question.
6. The CSV tab allows one to download the results in tabular form to your own computer.
7. From the SPARQL query tab, one can follow a link to the Yasgui service [34] and see the query used to retrieve the results from the LOD service.
8. The SHARE tab provides a permanent link to the search you have made. You can use it to find the search results later and to refer to the material.

People and Organizations Perspective In addition to letters, it is possible search and browse for individuals, families, groups and organizations, such as associations, clubs, and companies in the data in the PEOPLE AND ORGANIZATIONS view. This search view is useful, for example, when searching for and analyzing correspondences of a specific person or family. The main facet of the view is the name of the actor (Name), which is based on string search. The search criteria can be formed flexibly using, for ex-

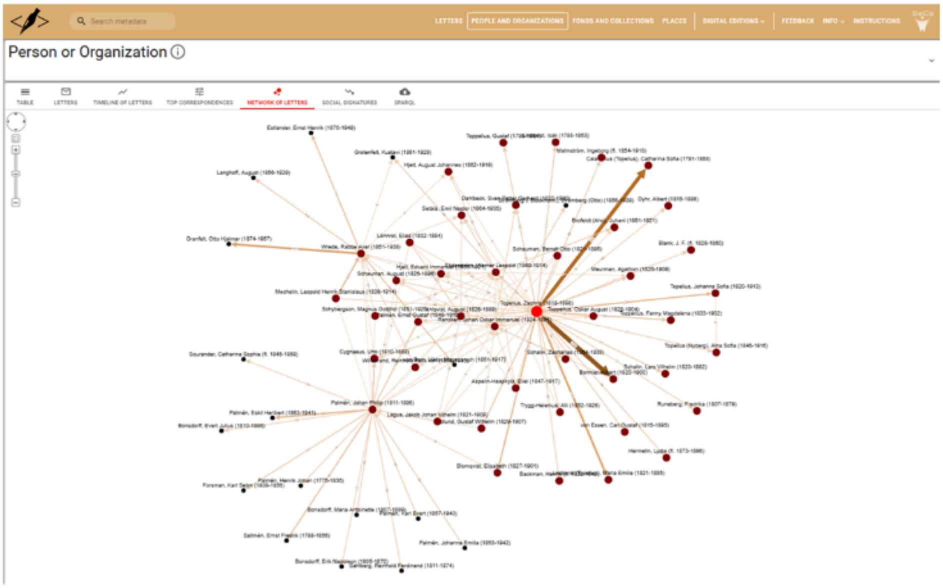


Figure 8. Instance page of Zachris Topelius. The the tab NETWORK OF LETTERS is chosen for visualizing the ego centric network of Topelius based on his correspondences. The nodes represent correspondents and arcs between them are letters. Wider arcs mean more letters. The network can be zoomed and browsed interactively.

ample, Boolean logic and the operators * and + as vague search symbols. More detailed instructions can be found via the facet's info button. There are 13 other facets available in the view.

The PEOPLE AND ORGANIZATIONS view allows one to access the target pages of individual actors, whose tabs include, for example, the General Information section (TABLE), which contains personal information in a tabular format. Here information about the actor can also be found pertaining to the collections in which the data about the actor is located. The actor's correspondences are compiled on the LETTERS tab, which also contains key quantitative measures used in network research (network metrics). There is also a separate tab for the timeline visualization of correspondences, TIMELINE OF LETTERS, and the visualization of the most active correspondents, TOP CORRESPONDENCES. Tab NETWORK OF LETTERS displays visually the correspondence network formed by the actors in the search result (cf. Fig. 8).

The identification of personal names, basic formatting (lemmatization) and linking to the portal's internal and external data sources, have been done automatically, and it is worth being prepared for errors in the material.

Fonds and Collections Perspective The FONDS AND COLLECTIONS perspective provides a faceted search for finding letters based on fonds and collections in which they are included.

If you are interested in the amount of correspondence or archivists in the collections of different organizations, this perspective provides a view for that. You can also view the collection level distributions of data through the facets of other views. The FONDS AND COLLECTIONS view tells you about the correspondence collections of different organizations and their archivists (fonds). You can also search and view archivists, based on their gender and profession.

Places Perspective Using the PLACES perspective, letters are visualized on a map based on their place of sending and receiving. By clicking on a marker, the links to the related letters shown in a pop-up window for further study. The PLACES view provides information about the places where letters were written, the places where they were received, and in some cases, the place names included in the content descriptions of the letters recorded in the metadata. For the digital editions data, where the textual contents of the letters are available, the letters can also be searched and analyzed using the places mentioned in them. In addition to the place descriptions included in the letter metadata, the PLACES view shows the location information necessary for forming a geographical hierarchy. For example, by selecting the category Italy, all places in Italy, such as Florence, will be included in the search. The subcategories with their hit counts can be seen in the hierarchical facets.

If the metadata of a person contains places related to his/her life cycle (such as places of birth and death) or the metadata of the letters contains location information, the actor can also be linked to data outside lsf, such as Biografiasampo, the National Library of Finland's Kanto register and Wikidata. The places are linked automatically, so the links should be treated with caution due to possible errors.

Digital Editions The portal includes four focused application perspectives for specific digital editions on the Web:

1. **Albert Edelfelt**. The view is based on the correspondence published online by the Swedish Literary Society (SLS): Albert Edelfelts brev[23].
2. **Elias Lönnrot**. The view is based on the collection Elias Lönnrotin kirjenvaihto[24], compiled and edited by the Finnish Literature Society.
3. **J. V. Snellman**. The view is based on the critical edition online publication J. V. Snellman Kootut teokset[25], managed by the Snellman Institute.
4. **Zachris Topelius**. The view shows Topelius' letters included in the edition Zacharias Topelius Skrifter[26] edited by SLS.

The LETTERSAMPO FINLAND perspectives offer alternative ways to search, analyze, and visualize previously published online letter data, which include not only metadata but also the contents of the letters. In contrast to the main application views, there are additional options to search and analyze letters using metadata about letter contents, such as the people or places mentioned in the letters, and get a link to the actual letter content in the digital edition.

The Snellman data contained rich manually made annotations and Finnish translations ready to use. Named entities were extracted using the Large Language Model (LLM)-based linker tool set of [39]. The Topelius letter metadata is currently used without additional content annotations, but is served with links to the primary letter contents. In the case of Albert Edelfelt's letters, human-made annotations about mentioned places and people were readily available in the metadata. For Elias Lönnrot letters this was not the case, and the letters were available in Swedish. As a remedy, the letter texts are being

[23] Albert Edelfelts brev: `https://edelfelt.sls.fi/`

[24] Elias Lönnrotin kirjenvaihto: `http://lonnrot.finlit.fi/`

[25] J. V. Snellman Kootut teokset: `https://snellman.kootutteokset.fi/`

[26] Zacharias Topelius Skrifter: `https://topelius.sls.fi/`

translated into Finnish by using the Llama 3.3 70B LLM [40]. Afterwards, the letters will be automatically annotated by using the aforementioned linking tool, which is implemented for Finnish texts. Formal evaluation of the annotation quality is still underway, but informal quality checks of the data suggest that the accuracy of the results is feasible and fit for the intended purpose.

7. Conclusions

The LETTERSAMPO system presented in this paper provides novel approaches to satisfy the information needs and related research questions set in Section 2. For example: 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 CH organizations in Finland, it seems that the 1.29 million letters harvested cover most of the Finnish letter data (1809–1917) available in professional archives, which is way more than was initially expected.

It was shown from a technical point of view, how to make distributed heterogeneous letter data FAIR for DH research and applications. The data harvesting and cleaning pipeline was tedious, required manual work, and the metadata available was in many cases incomplete and uncertain. However, the KG created was deemed useful, if the properties and limitations of the datasets are made transparent for the end user to support data literacy [41]. The LetterSampo Framework, including the Sampo model and Sampo-UI, used for the implementation worked well and were found re-usable.

It was also shown by examples, what kind of new historical insights can be obtained using DH methods on epistolary data. First studies have been conducted to provide new insights for research in humanities. The possibilities of using LETTERSAMPO in research are discussed in [42] and [43]. Network analyses using, e.g., the egocentric network based on the correspondences of the polymath Elias Lönnrot are reported in [19]. Article [21] 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 letter collection KG.

Evaluations regarding using faceted search and browsing, the basis of the Sampo UI model, suggest that this search paradigm is very usable when the user does not know exactly what (s)he is looking for [44,45]. Otherwise, traditional string based searching is usually preferred. To cater both needs at the same time, Sampo-UI has specific text search functionalities available, i.e, both search paradigms can be supported.

Further research on LETTERSAMPO includes more formal evaluation of the LOD publication and the portal from different perspectives, such as data quality, fitness for research use, and usability of the user interface of the portal. More research is already going on pertaining to textual semantic analyses of letters, including extracting and linking named entities, identifying keyword concepts, and topics of the letters.

Acknowledgments Thanks to Mikko Koho, Johanna Enqvist, Matti La Mela, and several research assistants for their 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 of the CSC – IT Center for Science were used in our work.

References

[1] Hotson H, Wallnig T, editors. Reassembling the Republic of Letters in the Digital Age. Göttingen University Press; 2019. Available from: `https://doi.org/10.17875/gup2019-1146`.

[2] van Miert D. What was the Republic of Letters? A brief introduction to a long history (1417–2008). Groniek. 2016;204/205:269-87.

[3] 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. 2022. In press. Available from: `https://arxiv.org/abs/2112.04336v1`.

[4] McCarty W. Humanities Computing. Palgrave, London; 2005.

[5] Gardiner E, Musto RG. The Digital Humanities: A Primer for Students and Scholars. New York, NY, USA: Cambridge University Press; 2015. `https://doi.org/10.1017/CBO9781139003865`.

[6] Dumont S. correspSearch – Connecting Scholarly Editions of Letters. Journal of the Text Encoding Initiative. 2016;(10).

[7] 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). CEUR Workshop Proceedings, vol. 2084; 2018. p. 76-88. Available from: `http://www.ceur-ws.org/Vol-2084/paper6.pdf`.

[8] Rockenberger A, Wiger EN, Witting MR, Bøe H, Thor EI, Wolden OJ, et al. Norwegian Correspondences and Linked Open Data. In: Navarretta C, Agirrezabal M, Maegaard B, editors. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference. vol. 2364 of CEUR Workshop Proceedings; 2019. p. 365-75. Available from: `http://ceur-ws.org/Vol-2364/33_paper.pdf`.

[9] Ahnert R, Ahnert SE, Coleman CN, Weingart SB. The Network Turn: Changing Perspectives in the Humanities. Elements in Publishing and Book Culture. Cambridge University Press; 2021. Available from: `https://doi.org/10.1017/9781108866804`.

[10] Burge C. The King's Gatekeeper: Thomas Cromwell, Epistolary Networks, and Power Structures at the Tudor Court, January—July 1540. Huntington Library Quarterly. 2023;86(2):257-81. Available from: `https://muse.jhu.edu/pub/56/article/936418`.

[11] Tuominen J, Koho M, Pikkanen I, Drobac S, Enqvist J, Hyvönen E, et al. 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. CEUR Workshop Proceedings, Vol. 3232; 2022. p. 415-23.

[12] 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. 2023;16(1).

[13] 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. Under open review. Available from: `https://seco.cs.aalto.fi/publications/2022/leskinen-et-al-lettersampo-2022.pdf`.

[14] Hyvönen E. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web. 2022. Accepted. Available from: `http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series`.

[15] Ikkala E, Hyvönen E, Rantala H, Koho M. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web. 2022;13(1):69-84.

[16] 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. CEUR Workshop Proceedings, Vol. 3508; 2023. Available from: `https://ceur-ws.org/Vol-3508/paper3.pdf`.

[17] Drobac S, Enqvist J, Leskinen P, Wahjoe MF, Rantala H, Koho M, et al. 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. University of Oslo Library, Norway; 2023. p. 248-62. Available from: `https://doi.org/10.5617/dhnbpub.10669`.

[18] Drobac S, Leskinen P, Wahjoe MF. Navigating the Challenges of Deduplicating Actors in Historical Letter Exchanges. In: Proceedings of the 24th European Conference on Knowledge Management. vol. 24. Academic Conferences International Limited; 2023. p. 1694-7. Available from: `https://doi.org/10.34190/eckm.24.2.1317`.

[19] 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; 2024. Under review. Available from: https://seco.cs.aalto.fi/publications/2024/poikkimaki-et-al-coco-2024.pdf.

[20] Hyvönen E, Leskinen P, Poikkimäki H, Rantala H, Tuominen J, Drobac S, et al. LetterSampo Finland (1809–1917) Data Service and Portal: Searching, Exploring, and Analyzing Historical Letters and Their Underlying Networks. In: Proceedings of ESWC 2025, supplement, poster and demo papers. Springer-Verlag; 2025. p. Accepted, forth-coming.

[21] La Mela M, Pikkanen I, Paloposki HL, (jouni tuominen@helsinki fi) JT. A Critical Collection History of Nineteenth-century Women's Letters. Overcoming the Occluded Archive with Data-Driven Methods. Digital Humanities Quarterly. 2025. Revised manuscript for DHQ Special Issue "Data Science and History: Practicing and Theorizing Data-Driven Inquiries into the Past".

[22] Isaac A. Europeana Data Model Primer. Europeana; 2023. Available from: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.

[23] Hyvönen E, Ahola A, Leskinen P, Rantala H, Tuominen J. How to Create a Portal for Digital Humanities Research Using a Linked Open Data Cloud of Cultural Heritage Knowledge Graphs: Case SampoSampo. In: Proceedings: SemDH 2025 Second International Workshop of Semantic Digital Humanities, co-located with ESWC 2025, Portoroz, Slovenia. CEUR Workshop Proceedings; 2025. Forth-coming. Available from: https://seco.cs.aalto.fi/publications/2025/hyvonen-et-al-samposampo-semdh-2025.pdf.

[24] Hickey TB, Toves JA. Managing Ambiguity In VIAF. DLib Magazine. 2014;20(7/8).

[25] Leskinen P, Hyvönen E. Biographical and Prosopographical Analyses of Finnish Academic People 1640–1899 Based on Linked Open Data. In: Proceedings of the Biographical Data in a Digital World 2022 (BD 2022), Tokyo. Institute of Cultural History, ZRC SAZU, Ljubljana, Slovenia; 2024. Available from: https://doi.org/10.3986/9789610508120_7.

[26] Leskinen P, Hyvönen E. Biographical and Prosopographical Analyses of Finnish Academic People 1640–1899 Based on Linked Open Data. In: Proceedings of the Biographical Data in a Digital World 2022 (BD 2022), Tokyo. Institute of Cultural History, ZRC SAZU, Ljubljana, Slovenia; 2024. Available from: https://doi.org/10.3986/9789610508120_7.

[27] Hyvönen E, Leskinen P, Tamper M, Rantala H, Ikkala E, Tuominen J, et al. BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In: The Semantic Web: ESWC 2019. Springer–Verlag; 2019. .

[28] Tamper M, Leskinen P, Hyvönen E, Valjus R, Keravuori K. Analyzing Biography Collection Historiographically as Linked Data: Case National Biography of Finland. Semantic Web. 2021. Forth-coming, preprint: https://seco.cs.aalto.fi/publications/2021/tamper-et-al-bs-2021.pdf.

[29] Annastiina Ahola EH Telma Peura. Using Linked Data for Data Analytic Literary Research: Case BookSampo - Finnish Fiction Literature on the Semantic Web. Journal of the Association for Information Science and Technology. 2025. Available from: https://doi.org/10.1002/asi.24984.

[30] Hyvönen E, Sinikallio L, Leskinen P, Drobac S, Leal R, Mela ML, et al. Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland. Semantic Web. 2024:Pre-Press, 1-23.

[31] Ahola A, Hyvönen E, Rantala H, Kauppala A. Historical Opera and Music Theatre Performances on the Semantic Web: OperaSampo 1830-1960. In: Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI. Proceedings of the 20th International Conference on Semantic Systems, 17-19 September 2024, Amsterdam, The Netherlands. IOS Press; 2024. p. 386-402.

[32] 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. Springer-Verlag; 2014. p. 226-30. Available from: https://doi.org/10.1007/978-3-319-11955-7_24.

[33] 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; 2024. Available from: https://ceur-ws.org/Vol-3759/paper4.pdf.

[34] Rietveld L, Hoekstra R. The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability. 2017;8(3):373-83.

[35] Li H. Social Network Extraction and Exploration of Historic Correspondences [PhD thesis]. Heidelberg University; 2018.

[36] Hyvönen E. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web. 2022;14(4):729-44.

[37] Ikkala E, Hyvönen E, Rantala H, Koho M. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web. 2022;13(1):69-84.

[38] 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. Available from: `https://ceur-ws.org/Vol-3508/paper3.pdf`.

[39] Leal R, Ahola A, Hyvönen E. Enriching Metadata with LLMs and Knowledge Graphs: Case Finnish Named Entity Linking. In: Digital Humanities in Nordic and Baltic Countries (DHNB 2025). Book of Abstracts; 2025. Long papers. Available from: `https://dhnb.eu/wp-content/uploads/2025/03/DHNB-2025_abstracts.pdf`.

[40] Llama Team AM. The Llama 3 Herd of Models; 2024.

[41] Koltay T. Data literacy for researchers and data librarians. Journal of Librarianship and Information Science. 2015;49(1):3-14.

[42] Enqvist J, Pikkanen I. Kirjeluettelot tutkimusaineistona ja kulttuuriperintönä: metadatan mahdollisuudet digitaalisen käänteen jälkeen. In: Karhu H, Kivilaakso K, Parente-Čapková V, editors. Tutkimuspolkuja yksityisarkistoihin – Aineistot historian, kulttuurin ja kirjallisuuden tutkimuksessa. Suomalaisen Kirjallisuuden Seura, hHElsinki; 2024. p. 390-426.

[43] Paloposki HL, Pikkanen I. Learning to Read Digital? Constellations of Correspondence Project and Humanist Perspectives on the Aggregated 19th-Century Finnish Letter Metadata. In: Baudry J, Burkart L, Joyeux-Prunel B, et al., editors. Digital History Switzerland 2024: Book of Abstracts; 2024. Available from: `https://digihistch24.github.io/submissions/444/`.

[44] Hearst M, Elliott A, English J, Sinha R, Swearingen K, Lee KP. Finding the flow in web site search. CACM. 2002;45(9):42-9.

[45] English J, Hearst M, Sinha R, Swearingen K, Lee KP. Flexible search and navigation using faceted metadata. University of Berkeley, School of Information Management and Systems; 2003.