LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data

EERO HYVÖNEN, Aalto University (SeCo) and University of Helsinki (HELDIG), Finland
PETRI LESKINEN, Aalto University (SeCo) and University of Helsinki (HELDIG), Finland
JOUNI TUOMINEN, University of Helsinki (HSSH and HELDIG) and Aalto University (SeCo), Finland

Epistolary data about historical letters is typically distributed in different archives depending on where the letters were sent to and received, and the data are represented using local heterogeneous data models and different natural languages. To study such letter data on a global level, the heterogeneous, distributed data in local siloes need to be aggregated and harmonized into larger services where local metadata can enrich each other to complement missing information. This paper presents a new framework, LetterSampo, for representing, publishing, and using epistolary data as Linked Open Data (LOD) on the Web for Digital Humanities (DH) research. The framework is used for creating LOD services and for building individual LetterSampo portals on top of them. To test and demonstrate the framework, it has been applied to the epistolary CKCC dataset of ca. 20 000 letters of the Huygens Institute, the Netherlands, to the correspSearch dataset of ca. 151 000 letters aggregated by the Berlin-Brandenburg Academy of Sciences and Humanities, and to the Early Modern Letters Online (EMLO) data of ca. 170 000 letters published by the University of Oxford. The CKCC and correspSearch datasets were published as LOD services, SPARQL endpoints, and as data dumps at Zenodo.org for re-use, and a demonstrational portal *LetterSampo: Historical Letters on the Semantic Web* was created based on this data. A novelty of the LetterSampo portals is to use faceted semantic search for filtering data of interest in flexible ways from multiple perspectives on two conceptual levels, and then visualize and analyze the results and data by seamlessly integrated data analytic tools—programming skills are not needed for using the portals. In addition to using the tools of the portal, the SPARQL endpoints can be used with modest knowledge about programming for DH research.

CCS Concepts: • **Information systems → Web searching and information discovery**; • **Applied computing → Arts and humanities**.

Additional Key Words and Phrases: Semantic Web, Linked Open Data, Data Analysis, Digital Humanities, Early Modern, Letter, Epistolary Data

ACM Reference Format:
Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. 2022. LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data. *ACM J. Comput. Cult. Herit.* x, x, Article x (July 2022), 23 pages. https://doi.org/X

1 INTRODUCTION

From Human Readable to Machine Understandable Epistolary Data

Epistolary data is distributed in different countries and collections, represented using different data models and formats, and written in different natural languages. To enable Digital Humanities (DH) research [6, 34] on heterogeneous, distributed letter collections, data about the letters have been aggregated, harmonized, and provided for the research

Authors' addresses: Eero Hyvönen, Aalto University (SeCo) and University of Helsinki (HELDIG), Finland, eero.hyvonen@aalto.fi; Petri Leskinen, Aalto University (SeCo) and University of Helsinki (HELDIG), Finland, petri.leskinen@aalto.fi; Jouni Tuominen, University of Helsinki (HSSH and HELDIG) and Aalto University (SeCo), Finland, jouni.tuominen@helsinki.fi.

Manuscript submitted to ACM

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].

The services of today are typically targeted for humans to read, and their contents are provided not as data for data-analytic DH research or for application development. This paper contributes to the state-of-the-art by applying the idea of Linked (Open) Data and Semantic Web technologies [9, 16] as the basis for creating epistolary databases and web services [45]: a new framework called *LetterSampo*[10] is presented and demonstrated for publishing and using epistolary data on the Web of Data for DH research.

LetterSampo Framework for Publishing and Using Epistolary Data

The key idea of the LetterSampo framework, focusing on representing epistolary metadata [52], is to refine the application domain-agnostic Sampo model [22] in order to create a domain-specific model and tool for publishing and studying epistolary (meta)data. The framework can be adapted and applied easily to creating new instances of data services and semantic portals based on different datasets that use a shared data infrastructure and principles of user interface (UI) design. A LetterSampo instance consists of a Linked Open Data (LOD) service and a semantic portal UI based on it.

It is argued in this paper that instead of creating a new custom UI for each dataset, the datasets can be adapted for a single UI model by using shared data models and principles for building portal UIs. On a data production level, this approach encourages content providers to create semantically interoperable data in their data silos, which is essential for aggregating epistolary collection data from distributed, heterogeneous, geographically distributed archives. From the portal usage viewpoint, more data can be provided to the end users via “standardized” UIs, which makes portals easier to learn to use. In this model, the idea of separating the underlying data service from the portal UI is encouraged: the data service can then be used for data analyses in DH research with little programming skills, and the portal can be used by ready-to-use tools for data search, exploration, and analysis.

To test and illustrate this approach, the LetterSampo framework has been designed, implemented, and applied to three datasets of historical letters with an emphasis on Early Modern times and the Republic of Letters (RofL) [14, 48, 49]:

(1) The CKCC corpus[11] is an aggregated collection of ca. 20 000 Dutch correspondences [48, 49].

(2) correspSearch is a metadataset of ca. 151 000 letters [3, 4] aggregated by the Berlin-Brandenburg Academy of Sciences and Humanities.

(3) Early Modern Letters Online (EMLO)[12] [13] is a database of ca. 170 000 letters aggregated and published by the University of Oxford. We transformed this database into Linked Data in a previous project [45], but as this dataset was not openly available as data due to copyright restrictions, this paper focuses on the CKCC and correspSearch datasets above.

[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] See project homepage http://seco.cs.aalto.fi/projects/rrl/ for detailed information and videos.
[11] CKCC is an acronym for “Circulation of Knowledge: A Web-based Humanities’ Collaboratory on Correspondences and Learned Practices in the 17th century Dutch Republic ”.
[12] http://emlo.bodleian.ox.ac.uk/home

In addition, the framework is currently being applied to 19th century epistolary datasets of the Grand Duchy of Finland in the Constellations of Correspondence (CoCo) project[13] [44].

As a result of the work reported in this paper, a LOD version of the CKCC corpus as well as of the correspSearch corpus have been created and made openly available for the research community in two ways:

(1) As SPARQL endpoints for CKCC[14] and correspSearch[15]. When the URL of either endpoint is used in a browser, the Yasgui[16] SPARQL editor [40] is automatically opened for querying and analyzing results interactively. Both datasets are hosted by the Linked Data Finland (LDF.fi) platform[17] [19].

(2) As data dumps at Zenodo.org for CKCC [24] and correspSearch [25].

We have also published a semantic portal demonstrator *LetterSampo: Historical Letters on the Semantic Web* that aggregates both CKCC and correspSearch data for everybody to test and use online[18]. The code of this demonstrator is available on GitHub[19]; it can be used as a template for implementing new portal instances of the LetterSampo framework.

The paper is organized as follows. Section 2 overviews the "Sampo model"[20] [22] for publishing and using LOD, and applies it to develop the LetterSampo framework for epistolary data. Next, the data models of LetterSampo are presented and data transformation discussed (Section 3). Section 4 presents from a user interface point of view what kind of portals can be created by using the LetterSampo framework and the Sampo-UI tool[21] [27] available for this purpose. After this (Section 5), the new opportunities of using the published LOD data and SPARQL endpoints DH in research are explored and challenges discussed. Finally, contributions of the paper are summarized, related works discussed, and deployment issues of the framework are considered (Section 6).

2 APPLICATION OF SAMPO MODEL TO LETTERSAMPO FRAMEWORK

This section introduces the Sampo model and applies it to create the LetterSampo framework.

Sampo Model Principles

Table 1. Sampo Model Principles P1–P6

P1	Support collaborative data creation and publishing
P2	Use a shared open ontology infrastructure
P3	Make clear distinction between the LOD service and the user interface (UI)
P4	Provide multiple perspectives to the same data
P5	Standardize portal usage by a simple filter-analyze two-step cycle
P6	Support data analysis and knowledge discovery in addition to data exploration

The Sampo model[22] [22] is a consolidated set of principles listed in Table 1 for collaborative publishing and using LOD on the Semantic Web. Principles P1–P3 lay a foundation for developing LOD services; principles P4–P6 are

[13] https://seco.cs.aalto.fi/projects/coco/
[14] See https://ldf.fi/ckcc/sparql
[15] See: https://ldf.fi/corresp/sparql
[16] https://yasgui.triply.cc
[17] https://ldf.fi
[18] https://lettersampo.demo.seco.cs.aalto.fi
[19] https://github.com/SemanticComputing/lettersampo-web-app
[20] https://seco.cs.aalto.fi/applications/sampo/
[21] Available in GitHub: https://github.com/SemanticComputing/sampo-ui
[22] The name "Sampo" comes from the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology.

related to creating semantic portals. The model is based on the Semantic Web standards[23] [12] and best practices of the W3C for Linked Data publishing [9, 16] and is supported by tools and infrastructures, such as the Sampo-UI framework for UI design. The model has evolved gradually in 2002–2022 when developing some twenty LOD services and portals that have had up to millions of end users, depending on the application[24]. Being domain-agnostic, the model has been used in a variety of application areas. For example: CultureSampo[25] aggregates and publishes a wide variety of tangible and intangible cultural heritage collections; the HealthFinland system[26] was used for publishing health promotion information; the Mapping Manuscript Migrations (MMM) system is an application[27] for pre-modern manuscript studies; in BiographySampo[28] and AcademySampo[29], biography and prosopography are in focus; NameSampo[30] is for toponomastic research on placenames; FindSampo[31] is used for studying archaeological finds; LawSampo[32] applies the model to publishing legislation and case law and ParliamentSampo[33] is for studying parliamentary speeches, political culture, and networks of politicians.

From a LOD service development point of view (P1–P3), the model is based on the idea of collaborative content creation (P1)[34]. The data is aggregated from local data silos into a global service, based on a shared ontology and publishing infrastructure (P2). The shared ontology infrastructure includes 1) shared *(meta)data models* for representing data (e.g., Dublin Core or CIDOC CRM) and 2) a set of *domain ontologies*[35] that are used for populating the instances of the data model classes, such as shared vocabularies for subject matter or historical places and actors. The local data are harmonized and enriched with each other by linking and reasoning. In this model everybody can arguably win, including the data publishers by enriched data and shared publishing infra, and the end users by richer global content and services. The model supports the idea of separating the underlying Linked Data service *completely* from the user interface via a SPARQL API (P3). This arguably simplifies the portal architecture and the data service can be opened for data analysis research in Digital Humanities. For example, the Yasgui interface for SPARQL querying and visualizing the results can be used, or Python scripting in Google Colab[36] and Jupyter notebooks[37].

Fig. 1 illustrates the principles P1–P3 in the context of creating a pan-European epistolary linked data collection and a data service as proposed in our earlier project [45]. In this case, epistolary data from different countries are aggregated by the EMLO service (see the directed red arcs in the figure) into a database and web service at the University of Oxford. This aggregated data is enriched mutually and transformed into a global linked data service (blue arc) with the idea that the data could then be accessed by the scholarly community using the linked data service SPARQL endpoint (dotted black arcs) for research and application development for portals. Unfortunately, opening the data could not be realized in this project due to copyright restrictions of the data.

[23] https://www.w3.org/standards/semanticweb/
[24] See https://seco.cs.aalto.fi/applications/sampo/ for more info, publications, and videos about the Sampo portals.
[25] Portal online: https://kulttuurisampo.fi; project homepage: https://seco.cs.aalto.fi/applications/kulttuurisampo/
[26] This prototype was deployed in public use in Finland by the National Institute for Health and Welfare; project homepage: https://seco.cs.aalto.fi/applications/tervesuomi/
[27] Portal online: https://mappingmanuscriptmigrations.org; project homepage: https://seco.cs.aalto.fi/projects/mmm/
[28] Portal online: https://biografiasampo.fi; project homepage: https://seco.cs.aalto.fi/projects/biografiasampo/en/
[29] Portal online: https://akatemiasampo.fi; project homepage: https://seco.cs.aalto.fi/projects/yo-matrikkelit/
[30] Portal online: https://nimisampo.fi; project homepage: https://seco.cs.aalto.fi/projects/nimisampo/en/
[31] Portal online: https://loytosampo.fi; project homepage: https://seco.cs.aalto.fi/projects/sualt/
[32] See LawSampo project homepage for more information and publications: https://seco.cs.aalto.fi/projects/lakisampo/
[33] See ParliamentSampo project homepage for more information and publications: https://seco.cs.aalto.fi/projects/semparl/
[34] In our case the collaborators are institutions rather than individual people.
[35] We use this term to refer to knowledge organization systems, such as SKOS vocabularies, that define terms and their relations in application domains.
[36] https://colab.research.google.com/notebooks/intro.ipynb
[37] https://jupyter.org

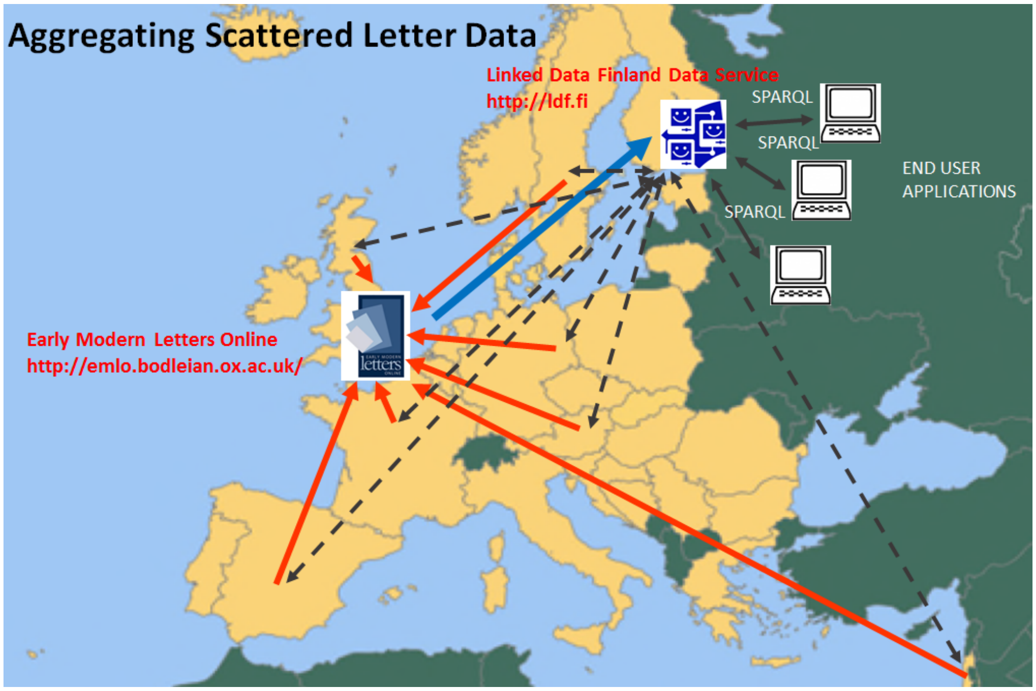

Aggregating Scattered Letter Data
Linked Data Finland Data Service
http://ldf.fi
SPARQL
SPARQL
SPARQL
END USER
APPLICATIONS
Early Modern Letters Online
http://emlo.bodleian.ox.ac.uk/

Fig. 1. Illustration of the Sampo model idea of using a Linked Data approach in aggregating and publishing distributed epistolary data based on the EMLO system [45]

In this paper, this idea is re-vitalized by using the open data sources of CKCC and correspSearch and by introducing the new LetterSampo framework for publishing and using epistolary data. In our case study, the aggregated datasets were already available, and the focus is on creating a semantic portal using principles P4–P6. They "standardize" the UI logic so that the portals are easier to use for the end users and for the programmers to develop [27]. Principle P4 articulates the idea of providing different thematic *application perspectives* by re-using the data service. The application perspectives can be provided on the landing page of the Sampo portal system or be completely separate applications by third parties. According to P5 the application perspectives can be used by a two-step cycle for research: First the focus of interest, the target group, is filtered using faceted semantic search [18, 43, 46]. Second, the target group is visualized or analyzed by using ready-to-use data analytic tools of the application perspectives. Finally, the Sampo model aims not only at data publishing with search and data exploration [33] but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analyzing, and even solving research problems in interactive ways (P6) [21]. Intelligent applications can be created more easily when the underlying data is based on well-defined machine-understandable semantics using the semantic web technology stack [12].

The Sampo model concerns only publishing and using data; it is assumed that there is a separate pipeline that creates the linked data in a harmonized form in a SPARQL endpoint based on the infrastructure. An example of such a pipeline for the WarSampo system (on World War II data) is presented in [30] and for the MMM Sampo (on medieval and Renaissance manuscript data) in [29]. The pipeline includes data transformations into a shared data model and aligning the entities mentioned by minting unique URI identifiers. In our case study demonstrator for the CKCC and correspSearch datasets, the entities were already aligned in both datasets with internal identifiers. During the transformation, the dataset-specific identifiers were aligned with each other by comparing the related data and by using new URIs for the entities.

LetterSampo Framework and Application Layers

The Sampo model principles have been used directly for creating semantic portals. However, its is also possible to apply them first to create an *application domain-specific framework* and reuse it for developing different related application instances in a particular application domain. Fig. 2 illustrates the idea on three conceptual levels:

(1) The highest (uppermost) conceptual level includes the Sampo model with its principles based on domain-agnostic, logical SW standards of the W3C and Linked Data publishing principles.

(2) The next level introduces domain-specific data models that can be used for populating metadata using domain-specific vocabularies and ontologies (e.g., concepts related to describing letters, historical persons and places, archives involved, etc.). This layer includes also a domain-specific template designed using the Sampo-UI framework that can be copied and used as a starting point for creating application instances. The template tells, e.g., what thematic application perspectives, data-analytic tools, and ready-to-use UI components are available in this application domain.

(3) Finally, applications can be created by adding in specific datasets into the framework, by creating a Sampo-UI implementation of the portal interface, and by publishing the data in a Linked Data service with a SPARQL endpoint. The figure depicts the LetterSampo framework with three applications corresponding to the datasets CKCC, EMLO, and correspSearch. By combining datasets based on shared ontologies, it is possible to create easily aggregated services; in our case the public LetterSampo demonstrator illustrates this by making use of both CKCC and correspSearch datasets.

Also the FindSampo framework [26] for archaeological find data is underway. In the case of FindSampo, archaeological find collections from the Finnish National Museum are used in one intance[38] and another one based on the Portable Antiquities Schema data of the British Museum is being developed using the same framework [37].

3 CREATING A LINKED DATA SERVICE FOR REPUBLIC OF LETTERS DATA

This section introduces first the application domain of our work, and then shows how the principles P1–P3 were used in creating a LOD service for epistolary data in our use case.

Republic of Letters

During the Age of Enlightment it became suddenly possible for people to send and receive letters across Europe and beyond, based on a revolution in postal services. This opportunity resulted into what the contemporaries called the *Respublica litteraria*, the Republic of Letters (RofL), a cross-national collaborative communication network that formed a basis for modern European scientific thinking, values, and institutions in Early Modern times 1400–1800. The CKCC corpus is a Dutch reflection of the RofL, a phenomenon that is in many ways analogous to phenomena on the contemporary Internet and World Wide Web (WWW) since the 1990's. For example, the Semantic Web on which this paper builds upon is a kind of ”Republic of Linked Open Data”, based on a yet another revolution in communication technologies and the underlying scientific community. From a network science point of view, the RofL and modern communications networks bear many interesting similarities [47].

[38] This application is in use on the Web at https://findsampo.fi.

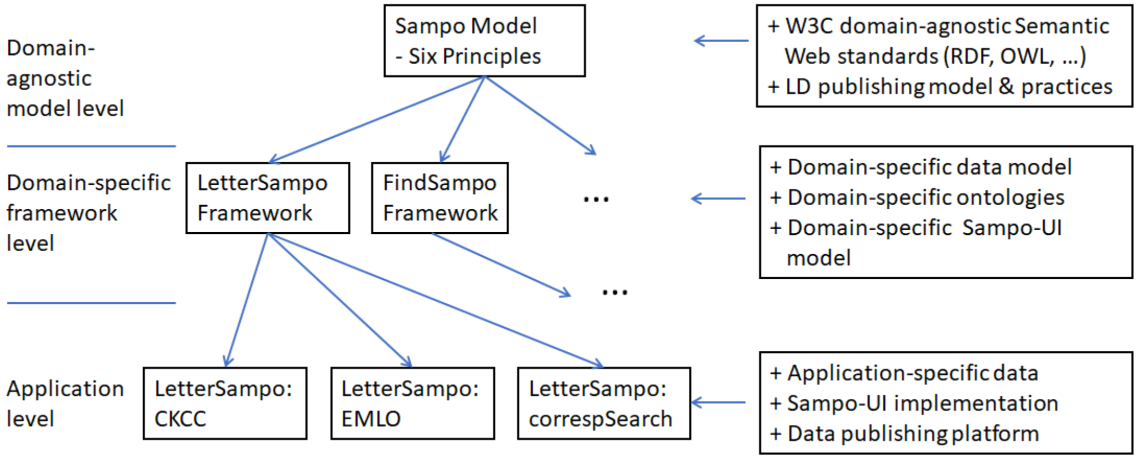

Domain-agnostic model level
Sampo Model
- Six Principles
+ W3C domain-agnostic Semantic Web standards (RDF, OWL, ...)
+ LD publishing model & practices
Domain-specific framework level
LetterSampo Framework
FindSampo Framework
...
+ Domain-specific data model
+ Domain-specific ontologies
+ Domain-specific Sampo-UI model
...
Application level
LetterSampo: CKCC
LetterSampo: EMLO
LetterSampo: correspSearch
+ Application-specific data
+ Sampo-UI implementation
+ Data publishing platform

Fig. 2. Three conceptual layers for creating Sampo portals: Sampo Model, Sampo frameworks, and applications. The idea is to re-use generic solutions of the model layer in domain-specific frameworks and then frameworks for application instances in different domains.

Correspondences of the RofL are typically archived in several countries and locations depending on where the letters were sent to. As a result, there is a great need for collaborative data creation when aggregating or reassembling the distributed letter data [14]. The CKCC corpus of RofL for our demonstrator aggregates data from the nine collections[39] listed in Table 2. The data was available by the open Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license from the Huygens Institute[40].

The CKCC contains data about over 20 000 letters. The data has been aggregated and harmonized earlier by the corpus developers into a database that was transformed into linked data in the LetterSampo project. The remaining task for us in creating the LOD service therefore was to design needed metadata models and create ontologies for populating the metadata element values.

Scholar	Gross # items	Net # items	Provider	Digitizer
Caspar van Baarle (Barlaeus)	505	505	UvA	Frans Blom, Marjolein van Zuylen
Isaac Beeckman	28	21	Huygens ING	Huib Zuidervaart
René Descartes	727	727	Utrecht University	Erik-Jan Bos
Hugo de Groot (Grotius)	8034	8034	Huygens ING	Henk Nellen
Christiaan Huygens	3090	3080	Huygens ING	Huib Zuidervaart
Constantijn Huygens	7297	7119	Huygens ING	Ineke Huysman
Antoni van Leeuwenhoek	282	282	Utrecht University / Huygens ING	Lodewijk Palm / Huib Zuidervaart
Dirck Rembrantsz van Nierop	80	80	Huygens ING	Marlise Rijks, Huib Zuidervaart
Jan Swammerdam	172	172	Huygens ING	Eric Jorink

Table 2. Datasets of the CKCC corpus used in the LetterSampo demonstrator

[39] http://ckcc.huygens.knaw.nl/?page_id=43
[40] http://ckcc.huygens.knaw.nl

The correspSearch dataset has been aggregated from a large network of content providers[41] using an XML-based Correspondence Metadata Interchange (CMI) Format, based on the Text Encoding Initiative (TEI)[42], for harvesting and harmonizing the data.

Table 3. RDF schema for Letter, Actor, and Place classes. Column *C* marks the cardinality of the property.

Property URI	**C**	**Range**	**Meaning of the value**
LETTER (:Letter, subclass of crm:E84_Information_Carrier and crm:E33_Linguistic_Object			
skos:prefLabel	1	xsd:string	Preferable label
:was_addressed_to	0..1	crm:E39_Actor	Recipient of the letter
:was_sent_from	0..1	crm:E53_Place	Place of sending
:was_sent_to	0..1	crm:E53_Place	Place of receiving
:has_time	0..1	crm:E52_Time-Span	Time of sending
dct:subject	0..n	skos:Concept	Subject matter of a letter (not systematically available in our case study data)
:source	1..n	rdfs:Resource	Used data source
:in_tie	1	:Tie	Correspondence in which this letter belongs to
ACTOR (crm:E21_Person)			
skos:prefLabel	1	xsd:string	Preferable label
:created	0..n	:Letter	Created letter
:birthDate	0..1	crm:E52_Time-Span	Time of birth
:birthPlace	0..1	crm:E53_Place	Place of birth
:flourished	0..1	crm:E52_Time-Span	Time of flourishing
:deathDate	0..1	crm:E52_Time-Span	Time of death
:deathPlace	0..1	crm:E53_Place	Place of death
:has_statistic	1...n	:NetworkStatistic	Precalculated network statistics, e.g., centrality measures
:source	1..n	rdfs:Resource	Used data source
PLACE (crm:E53_Place)			
crm:P89_falls_within	0..1	crm:E53_Place	Place higher in hierarchy
skos:prefLabel	1	xsd:string	Preferable label
geo:lat	0..1	xsd:decimal	Latitude of the coordinates
geo:long	0..1	xsd:decimal	Longitude of the coordinates

Data Model and Ontologies

The shared ontology infrastructure in LetterSampo contains 1) a metadata model and 2) a set of domain ontologies that are used for populating the model. The metadata model includes three core classes: Actor, Letter, and Place. Their instances are searched for and analyzed in a LetterSampo LOD service and portal.

Table 3 presents the metadata schemas for the three core classes using the namespaces *skos*[43] (SKOS Simple Knowledge Organization System), *xsd*[44] (XML Schema), *crm*[45] (Erlangen (CIDOC) CRM), *dct*[46] (Dublin Core DCMI Metadata Terms), *geo*[47] (W3C Basic Geo Vocabulary), and the default LetterSampo schema domain namespace *http://ldf.fi/schema/lssc/*. During the data transformation, also instances of CIDOC CRM timespans were created, as well as instances of directed connections between correspondents indicating the number of letters sent between the actors[48]. This is an example of enriching the data with simple reasoning. The original CKCC data was available as linked data from our earlier project on EMLO data; the correspSearch data was available in tabular CSV form that was transformed to conform to the model of Table 2. This model was selected because it matched directly to the datasets of

[41] https://correspsearch.net/en/data.html
[42] https://tei-c.org
[43] http://www.w3.org/2004/02/skos/core#
[44] http://www.w3.org/2001/XMLSchema
[45] http://erlangen-crm.org/current-version
[46] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[47] https://www.w3.org/2003/01/geo/
[48] The data model used makes use of some CIDOC CRM constructs but does not follow the standard CIDOC CRM practice.

CKCC, correspSearch, and EMLO. Furthermore, the three core classes could be used in a natural way as a basis for the application perspectives for creating the portal using the Sampo-UI tool. The idea of application perspectives will be elaborated in more detail in the next section.

Data Transformation

The key ontologies for Letters, Actors, and Places were generated in a bottom-up fashion by creating new instances of the three classes as needed based on the metadata element values used in the original data. This simple approach was enough for creating linked data within the dataset, i.e., for "four star" linked data according to the 5-star deployment scheme of Tim Berners-Lee[49]. In this case, shared pre-defined ontologies were not used because encompassing ontologies were not available for this application domain. However, the data was enriched by using additional ontology alignments to shared ontologies. For example, Wikidata was used for this purpose to enrich, e.g., the place instances with geocoding and the biographical data about the correspondents.

Fig. 3 depicts an extract of the RDF metadata for a letter, in this case one sent by Gottfried Leibniz to Christiaan Huygens. According to the data model, the letters are instances of the class *:Letter*, actors of *crm:E21_Person*, and places of *crm:E53_Place*. In the example instance (*letters:l910593*), the recipient (*actors:p11763*) is represented with the property *:was_addressed_to*, and the places of sending (*places:p900116*) and receiving (*places:p300017*) with the properties *was_sent_from* and *was_sent_to*, respectively. In the example only some of the properties for actors, letters, and places are shown while more detailed information, e.g., place coordinates and birth information of an actor are left out for brevity. The resource for the sender has the property *:created* connecting the sender to a letter. The property *:has_time* indicates the time of sending. The property *:source* relates the letter to the source data, in this case the data collection of Christiaan Huygens (Table 2). The property *:is_related_to* provides a link to a page in ePistolarium[50] depicting, e.g., the full text of that particular letter. The property *:in_tie* binds the letter resource to a *:Tie* object *ties:p11763-p13152*. The *:Tie* objects model the entire correspondence between the two actors, and they where constructed from the source data to facilitate network research as well as to ease the queries of the web portal. Each *:Tie* object contains the URIs for both the actors involved (*:actor1*, *:actor2*) as well as the precalculated number of letters in the correspondence (*:num_letters*).

The principle P3 recommends to "make clear distinction between the LOD service and the user interface (UI)". The motivation behind this principle is to make the re-use of the data easier for research using different kind of tools, and to support developers in making new applications on top of the data service. Arguably, separating the data service from the application logic and UI also makes development work more modular and easier, as suggested in many application development frameworks, such as the Model-View-Controller (MVC) model[51].

On top of a SPARQL API it is possible to define more dedicated and simpler APIs depending on the application and user needs, such as grlc [35]. SPARQL querying can be computationally inefficient and has lead some developers to use other tools for developing search engines, such as Elasticsearch[52] and Solr[53]. However, SPARQL servers, such as Fuseki[54] used in our LetterSampo demonstrator portal, include efficient tooling for text indexing, such as Lucene[55] and Solr.

[49] https://5stardata.info/en/
[50] http://ckcc.huygens.knaw.nl/epistolarium/letter.html?id=huyg003/2206
[51] https://www.brainvire.com/six-benefits-of-using-mvc-model-for-effective-web-application-development/
[52] https://www.elastic.co/elasticsearch/
[53] https://solr.apache.org
[54] https://jena.apache.org/documentation/fuseki2/
[55] https://lucene.apache.org

```
@prefix : <http://ldf.fi/schema/lssc/> .
@prefix actors: <http://ldf.fi/lssc/actors/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dct: <http://purl.org/dc/terms/>
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix letters: <http://ldf.fi/lssc/letters/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix places: <http://ldf.fi/lssc/places/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

letters:l910593 a :Letter ;
    :was_addressed_to actors:p11763 ;
    :was_sent_from places:p900116 ;
    :was_sent_to places:p300017 ;
    :has_time times:time_1679-12-10T00:00:00-1679-12-10T23:59:59-1679-12-10T00:00:00-1679-12-10T23
     :59:59 ;
    :in_tie ties:p11763-p13152 ;
    :is_related_to <http://ckcc.huygens.knaw.nl/epistolarium/letter.html?id=huyg003/2206> ;
    :source <http://ldf.fi/ckcc/sources/source_Huygens%2C+Christiaan> ;
    skos:prefLabel "10 Dec 1679: Leibniz, Gottfried Wilhelm, 1646-1716 (Hannover, Lower Saxony,
     Germany) to Huygens, Christiaan, 1629-1695 (Paris, Île-de-France, France)"@en .

actors:p13152 a crm:E21_Person ;
    :created letters:l910593 ;
    skos:prefLabel "Leibniz, Gottfried Wilhelm, 1646-1716" .

actors:p11763 a crm:E21_Person ;
    skos:prefLabel "Huygens, Christiaan, 1629-1695" .

places:p900116 a crm:E53_Place ;
    skos:prefLabel "Hanover" .

places:p300017 a crm:E53_Place ;
    skos:prefLabel "Paris" .

ties:p11763-p13152 a :Tie ;
    :actor1 actors:p11763 ;
    :actor2 actors:p13152 ;
    :num_letters 70 ;
    :has_time times:time_1674-01-01T00:00:00-1695-12-31T23:59:59-1674-01-01T00:00:00-1695-12-31T23
     :59:59 ;
    skos:prefLabel "Huygens, Christiaan <-> Leibniz, Gottfried Wilhelm"@en .
```

Fig. 3. Extract of the RDF data for a letter from Gottfried Wilhelm Leibniz (1646–1716) to Christiaan Huygens (1629–1695)

In Sampo systems, the transformed data is published according to the Linked Data principles [9], including, e.g., content negotiation for dereferencing HTTP URIs and, most importantly, a SPARQL endpoint with an API. In our work, the Linked Data Finland platform [19] has been used for hosting the data but any other LOD platform could be used for creating the needed SPARQL endpoints in the same way. If the data in two endpoints conform to the same LetterSampo framework data model, the same UI can be used in both of them by just switching the endpoint address with little adaptation.

4 USER INTERFACE MODEL FOR SEMANTIC SAMPO PORTALS

This section explains how the principles P4–P6 are used in the LetterSampo framework for creating portal user interfaces with Sampo-UI. The focus is on describing what kind of UIs can be created from the application end user's perspective (not from a technical programming point of view).

Supporting Multiple Application Perspectives on Two Levels

A key goal of the Sampo and Sampo-UI model is to facilitate research about cultural entities, such as letters, and groups of them:

(1) *Group level*. DH research is often focused on groups of entities that share some common characteristics. In LetterSampo, for example, groups of interest may include letters written by a person or groups of persons active in a certain time period, or groups of correspondents (people) born in a particular area and/or belonging to the same social or professional class.

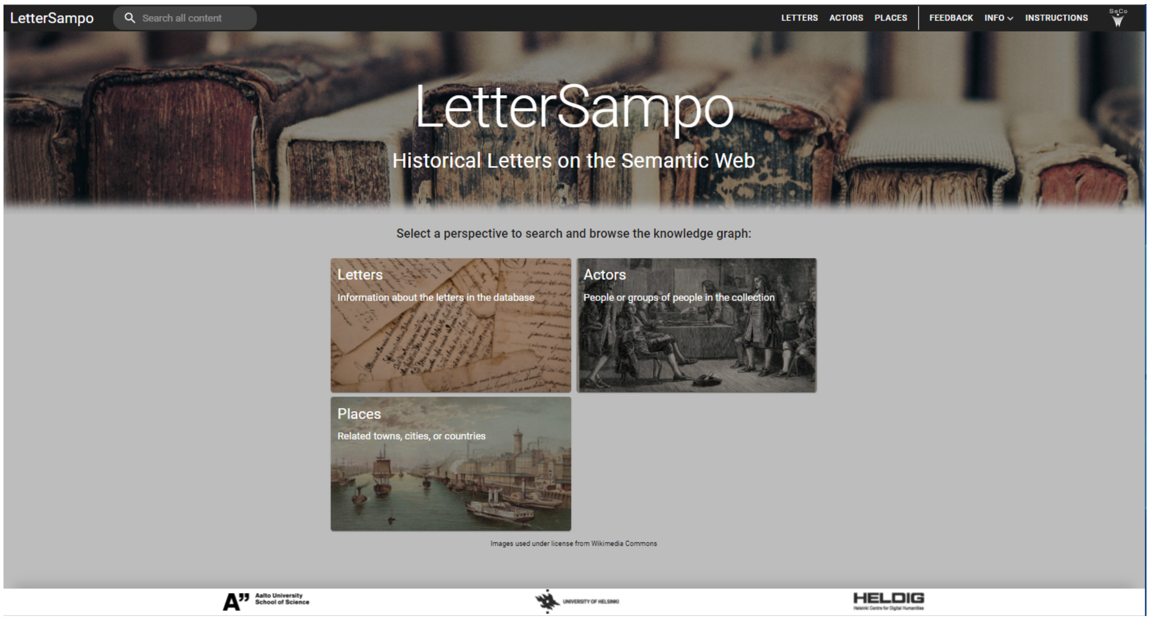
(2) *Entity level*. On the entity level, particular members of groups are studied in context. For example, a letter entity can be studied in term of its data content and metadata by close reading, in relation to related correspondences, and a particular person entity can be investigated in relation to his/her social networks, to letters (s)he sent and received, or places (s)he visited.

This idea of two levels originates from our work in the biographical domain, especially when developing the system "BiographySampo – Finnish Biographies on the Semantic Web" [17], where both biographical research of particular biographees is supported as well as prosopographical research on groups of people using data-analytic tools and visualizations [42]. In Sampo-UI-based systems, groups are subsets of selected core classes in the underlying LOD service and entities are instances of these classes. In the case of LetterSampo, the core classes are Letters, Actors, and Places as described in Table 3.

The Sampo model principle P4 argues for "providing multiple perspectives to the same data". In the case of epistolary data there is a need for searching, exploring, and analyzing the (meta)data from the perspectives of the core classes for letters, corresponding people, and historical places. When using Sampo-UI, the selected perspectives are provided on the landing page of the portal depicted in Fig. 4 for our demonstrator "LetterSampo – Historical Letters on the Semantic Web"[56]. In the Letters perspective, letters can be filtered, searched, and studied; the Actors perspective provides a view to the letter data based on the communicating people; in the Places view, the idea is have a look at the data based on locations on maps illustrating, e.g., where the letters were sent and received, and where the correspondents were therefore situated in different times.

The Sampo model principle P5 suggests to "standardize portal usage by a simple filter-analyze two-step cycle". To accommodate this principle on the group level, the group level services are provided as *application perspective pages*, one for each class of entities. These pages have faceted search engines for filtering and studying groups of entities (instances of the classes Letter, Actor, and Place), and for applying data analyses and visualizations to the filtered result groups. For studies on the entity level, each entity has a landing page of its own, its *entity homepage*. It aggregates data and links related to the entity and provides tools for data analyses of the entity in its linked data context. Links to the entity homepages can be found on the application perspective pages for the filtered entities.

[56] Available at: https://lettersampo.demo.seco.cs.aalto.fi



Select a perspective to search and browse the knowledge graph:

Information about the letters in the database

People or groups of people in the collection

Related towns, cities, or countries

Images used under license from Wikimedia Commons

Fig. 4. LetterSampo landing page with three application perspectives, based on CKCC and correspSearch datasets

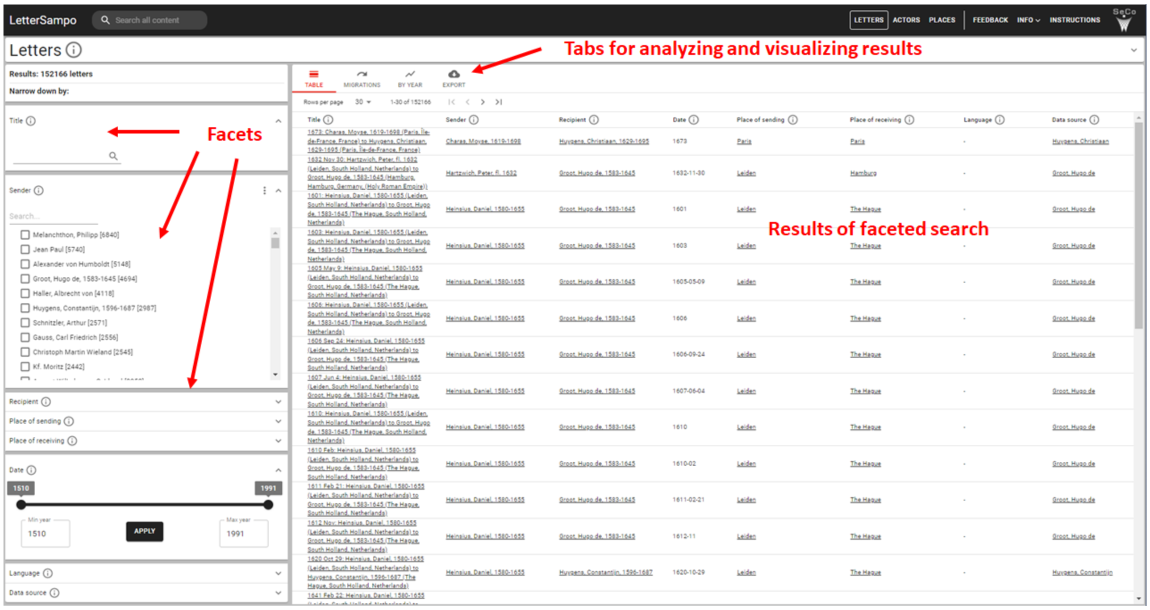
In the following it is illustrated, how the three application perspectives of the LetterSampo framework on both group and entity levels can be used for studying epistolary data. For this purpose, examples using the demonstrator “LetterSampo – Historical Letters on the Semantic Web” based on the CKCC and correspSearch datasets are presented.

Letters Perspective

By clicking on ”Letters” on the landing page (Fig. 4), the corresponding faceted search application perspective page is opened with facets Sender, Recipient, Place of sending, Place of receiving, Date, Language, and Data source (Fig. 5). By opening the facets and by making selections on them, letter groups can be filtered. In the figure, the Sender and Date facets are opened. By clicking on *Gauss, Carl Friedrich [2556]* on the Sender facet, the 2556 letters sent by Gauss would be filtered. In the same way, the Date facet value range can be modified for focusing on a certain time period of letters. The results are shown on the right with paging.

The filtered letter groups can be studied by using the four tabs over the results’ pane:

(1) TABLE tab is the default visualization and shows the results as a table. The results can be sorted by clicking the column labels.
(2) MIGRATIONS tab shows the routes of letters on a map as arcs from the place of sending (blue color end) to place of receiving (red color end), as illustrated in Fig. 6.
(3) BY YEAR tab shows the number of letters on a timeline based on Date values.
(4) EXPORT tab can be used to view the data query on the Yasgui SPARQL editor that can be used to download the data and as an easy access to query the data further using SPARQL.



Results: 152166 letters
Narrow down by:

Title

Sender

Search...

- □ Melanchthon, Philipp [6840]
- □ Jean Paul [5740]
- □ Alexander von Humboldt [5148]
- □ Groot, Hugo de, 1583-1645 [4694]
- □ Haller, Albrecht von [4118]
- □ Huygens, Constantijn, 1596-1687 [2987]
- □ Schnitzler, Arthur [2571]
- □ Gauss, Carl Friedrich [2556]
- □ Christoph Martin Wieland [2545]
- □ Kf. Moritz [2442]

Recipient

Place of sending

Place of receiving

Date

1510 – 1991

Min year 1510

Max year 1991

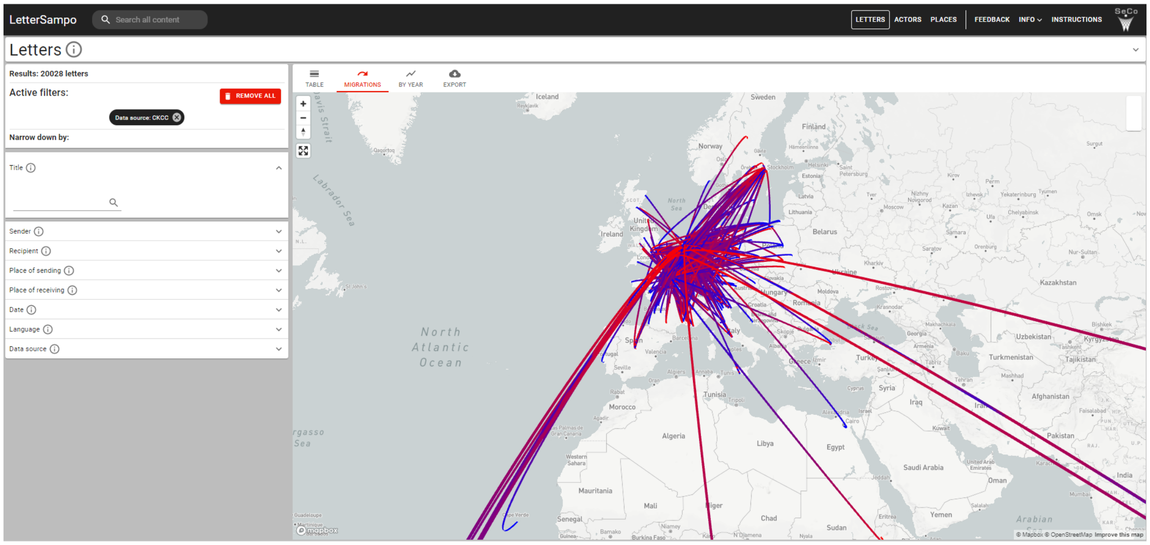
Language

Data source

Rows per page 30 1-30 of 152166

Title	Sender	Recipient	Date	Place of sending	Place of receiving	Language	Data source
1673: Charas, Moyse, 1619-1698 (Paris, Île-de-France, France) to Huygens, Christiaan, 1629-1695 (Paris, Île-de-France, France)	Charas, Moyse, 1619-1698	Huygens, Christiaan, 1629-1695	1673	Paris	Paris	-	Huygens, Christiaan
1632 Nov 30: Hartzwich, Peter, fl. 1632 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (Hamburg, Hamburg, Germany, (Holy Roman Empire))	Hartzwich, Peter, fl. 1632	Groot, Hugo de, 1583-1645	1632-11-30	Leiden	Hamburg	-	Groot, Hugo de
1601: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1601	Leiden	The Hague	-	Groot, Hugo de
1603: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1603	Leiden	The Hague	-	Groot, Hugo de
1605 May 9: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1605-05-09	Leiden	The Hague	-	Groot, Hugo de
1606: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1606	Leiden	The Hague	-	Groot, Hugo de
1606 Sep 24: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1606-09-24	Leiden	The Hague	-	Groot, Hugo de
1607 Jun 4: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1607-06-04	Leiden	The Hague	-	Groot, Hugo de
1610: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1610	Leiden	The Hague	-	Groot, Hugo de
1610 Feb: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1610-02	Leiden	The Hague	-	Groot, Hugo de
1611 Feb 21: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1611-02-21	Leiden	The Hague	-	Groot, Hugo de
1612 Nov: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Groot, Hugo de, 1583-1645 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Groot, Hugo de, 1583-1645	1612-11	Leiden	The Hague	-	Groot, Hugo de
1620 Oct 29: Heinsius, Daniel, 1580-1655 (Leiden, South Holland, Netherlands) to Huygens, Constantijn, 1596-1687 (The Hague, South Holland, Netherlands)	Heinsius, Daniel, 1580-1655	Huygens, Constantijn, 1596-1687	1620-10-29	Leiden	The Hague	-	Huygens, Constantijn
1641 Feb 22: Heinsius, Daniel, 1580-1655							

Fig. 5. Letters application perspective page for searching and studying letters



Results: 20028 letters

Active filters:

Narrow down by:

Title

Sender

Recipient

Place of sending

Place of receiving

Date

Language

Data source

Fig. 6. Migrations of 20 028 letters send and received using the whole CKCC data corpus, visualized on the MIGRATIONS tab. The data was filtered by selecting category CKCC on the Data source facet.

By clicking on a letter link in the results table, the corresponding letter entity page is opened with data about the letter and links to related entity pages of people and places. In the same way as in the perspective pages, an entity page may have tabs for studying and visualizing the entity in different ways, e.g., for showing the actual letter if this is available in the data. New tabs can be created as needed.

The entity pages can also be linked to additional *information pages* that have been created based on the data. To illustrate this, the demonstrator includes information pages for each correspondence connection between two people. This information page provides the user with tabs for studying the letter in the context of the correspondences between a sender and a receiver, including the following:

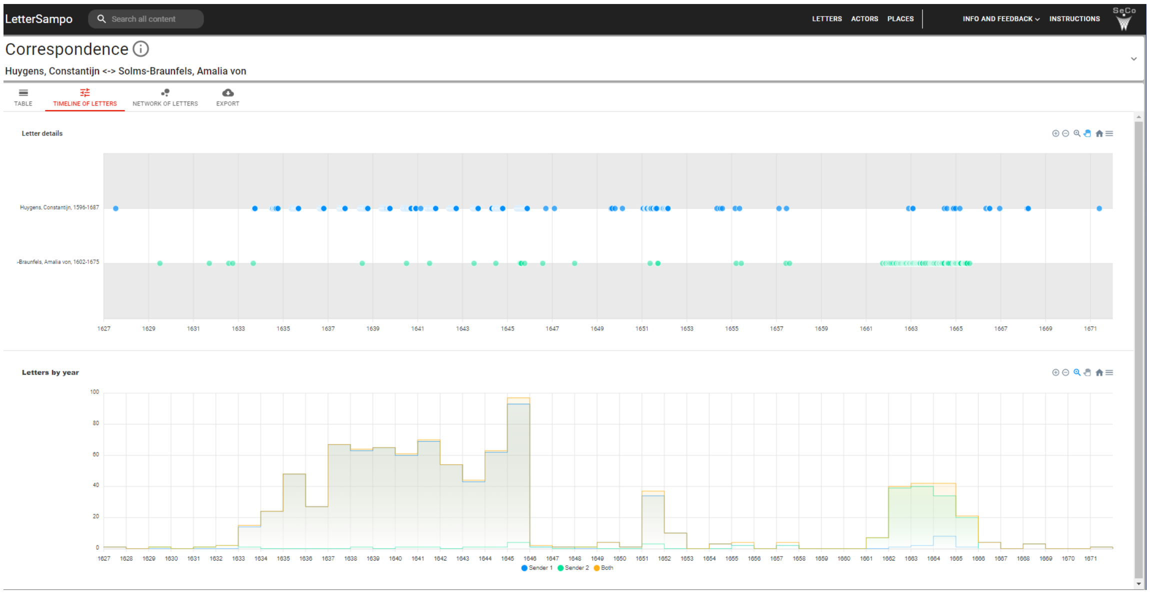


Fig. 7. TIMELINE OF LETTERS tab visualization of correspondences between Contstantijn Huygens older and Amalia von Solms-Braunfels. The upper timeline shows letters sent by Contstantijn Huygens as blue spots, the lower graph letters of Amalia von Solms-Braunfels as green spots. The histogram below depicts the yearly number of letters exchanged.

(1) TABLE view includes detailed information about the correspondences between the two specific actors.
(2) TIMELINE OF LETTERS details data about all the letters sent and received between the two actors (cf. Fig. 7 for an example). Two timeline charts on this tab show the correspondence with precision of a single day or by years. One can zoom or pan the timelines by using the buttons on the visualization components.
(3) NETWORK OF LETTERS contains a network visualization showing the closest other correspondences connecting to the specific correspondence between the two actors of this page.

Actors Perspective

By clicking the icon ”Actors” on the landing page (Fig. 4), a faceted search application perspective is opened for filtering and finding actor groups by their name, gender, (actor) type, and times of birth or death. Faceted search can be used by making category selections of the facets, e.g., Birth time = 1650–1700 and Gender = Female. After each selection, a hit count is calculated for all categories indicating the number of search results if that category is selected next. This helps the users in making next search selections and the process never ends up in a dead end with no hits. In Sampo-UI, the hit count distributions along the facet dimensions can also be used for statistical analyses and visualizations.

Faceted search can also be used for traditional text search by using a text search facet. This is useful when the user knows what (s)he is looking for and is able to type in, e.g., the name of the person searched for [5]. For example, we can search with "Huygens" and find in the result list lots of members of the Huygens family, such as Constantijn Huygens (older) (1596–1687), the poet and statesman, his son Constantijn Huygens (younger) (1628–1697), the brother of the Dutch physicist, mathematician, astronomer, and inventor Christiaan Huygens (1629–1695).

By clicking on an actor link on the result list, the homepage of the actor is opened. The entity homepage of an actor contains related metadata and includes several tabs for analyzing the activities of the actor:

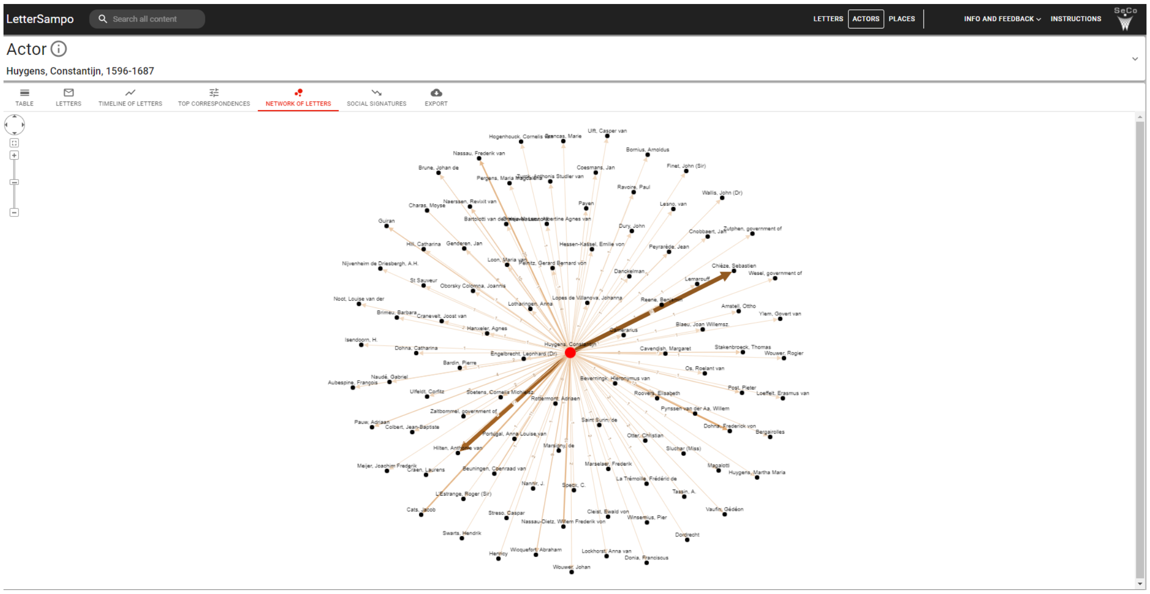


Fig. 8. Ego-centric network of letters of Constantijn Huygens (1596–1687), the poet, based on his correspondences in the CKCC data. The dots represent people, Constantijn Huygens marked in red, and arcs between them illustrate letters sent. The width of the arc indicates the number of letters exchanged.

(1) TABLE tab shows the basic biographical information of the actor including his/her alternative names and vocations, and the letters sent and received, based on the underlying metadata. There is also a list of all the other actors with whom (s)he has been in contact with, and recommendation links to additional information pages.

(2) LETTERS tab show the available information about the letters sent or received by the person, if such data is available in the knowledge graph used.

(3) TIMELINE OF LETTERS tab is used to visualize the number of letters on a timeline. Fig. 9 depicts the timeline of Constantijn Huygens' (1596–1687) correspondences.

(4) TOP CORRESPONDENCES view shows a timeline of letters and most important correspondences of the person. The upper chart of the timeline shows the activities using the precision of one day, the lower one the yearly amounts of sent and received letters. The user can zoom or pan the timelines by using the buttons of the visualization components.

(5) NETWORK OF LETTERS tab visualizes the correspondences as an egocentric network based on correspondences for network analysis [7, 38]. Fig. 8 depicts the network of Constantijn Huygens (1596–1687).

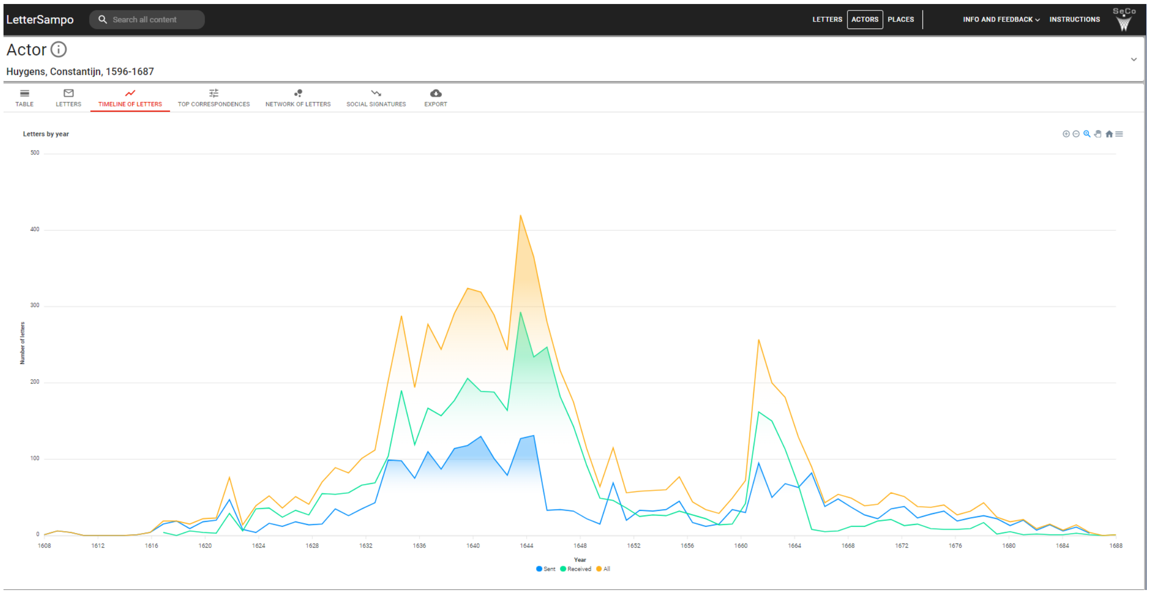


Fig. 9. Timeline of letters sent (lower line in blue), received (middle line in green) by Constantijn Huygens (1596–1687), and in total (upper line in yellow) based on sent and received letters in the underlying dataset

(6) SOCIAL SIGNATURES view has a chart showing how much the actor has been in correspondence with the most, the second most, etc, important other actors during the different time periods along the total time of activity [10].

(7) EXPORT tab shows the SPARQL query used to generate the result table in the Yasgui query editor. The tab EXPORT can be used for downloading the data in, e.g., CSV form for external data analysis applications, such as spreadsheet programs and R[57] for statistical computing.

Places Perspective

The third main application perspective ”Places” can be used on the group level for finding places of interest. As in the other perspectives, each place has an entity homepage on which linked data related to the place is aggregated to support studying and exploring the data from a geographical perspective.

It should be emphasized that any visualization or data analysis in a system like LetterSampo is an analysis of the underlying dataset that is only an incomplete reflection of the underlying real world. Any conclusions drawn from the analysis needs source criticism and understanding the limitations of the data. For example, in the LetterSampo demonstrator the underlying metadata probably misses many letters sent between the Huygens family members. In spite of this, analyzing an incomplete letter collection by distant reading can give novel insights to the researcher and collection manager, and suggest topics of interest for further close reading and study.

In addition to missing data about letters altogether, another form of incompleteness in the datasets is that in some cases a metadata element value for an instance, say the place where the letter was received, is missing. In Sampo-UI this problem can be made transparent to the end user by a special facet category "Unknown" with a hit count of its own.

[57] https://www.r-project.org/

5 USING THE LOD SERVICE FOR RESEARCH

The open SPARQL endpoints and the data underlying Sampo portals can be used for DH research not only via the portal UI but also via APIs. This section explores these possibilities for research and application development in the context of the LetterSampo framework.

Exporting Data from the SPARQL Endpoint for Data Analyses

An obvious way of reusing the data service is to download data, transfer the data into whatever format is needed for the tool to be used, such as R for statistical analyses. An example of using this approach is presented in [47]. In this case study, the linked data publications of CKCC and correspSearch data were re-used for network analysis, using custom made tools created earlier in other research projects. Here the RofL data was analysed and compared with four modern datasets of mobile phone call networks, emails, community boards, and wall-postings on a social media platform. The analyses indicated striking resemblances between contemporary and epistolary communication network patterns.

A particular benefit of using the LOD repository for data export is that by using SPARQL its is easy to filter from a large dataset more focused subsets of interest, both in terms of size and structure of the export. In the latter case, SPARQL CONSTRUCT queries are a handy tool for restructuring data. The result sets created using a SPARQL endpoint are easily available in different formats for different use cases, such as in XML and in tabular format. For exporting data, the Sampo-UI tool also has a generic component by which data exports can be created on a separate tab as discussed in Section 4.

Using the SPARQL Endpoint for Customized Analyses

The underlying SPARQL endpoint can be used for custom data analyses using, e.g., the Yasgui query editor and Python scripting with Google Colab and Jupyter notebooks. This is a way to provide the aggregated linked data openly back to the data providers and the scientific community in the large for their own scholarly work. This is an extension of services with respect to traditional portals, such as EMLO, ePistolarium, and correspSearch, where the aggregated data is provided back, too, but only as human readable web pages for close reading and not as a data service for machines to use.

For example, Fig. 10 visualizes the correspondence activities of the mathematicians in the CKCC data, based on their mutual ranking (y-axis) in 1640–1690. This analysis and visualization were created using a Python script in a Google Colab notebook[58] that first makes a SPARQL query to the data service filtering data of interest about the mathematicians. The query result can then re-formatted and visualized using the Python data analysis libraries NumPy[59], SciPy[60], Pandas[61], and Matplotlib[62] or the Javascript library ApexCharts[63] for visualization. In the same vein, Fig. 11 illustrates the ranking of places mentioned in the data (the number of placename occurrences), aggregated by decades, on a timeline.

Analyses such as these require some computational skills on using SPARQL and Python but can be very useful for researchers in humanities as suggested and exemplified in [1] using the MMM LOD service on medieval manuscripts. In this way, the data providers can access flexibly their own data without the constraints of their cataloging systems and collection publishing applications.

[58] Notebook available at https://doi.org/10.5281/zenodo.6122979
[59] https://numpy.org
[60] https://www.scipy.org
[61] https://pandas.pydata.org
[62] https://matplotlib.org
[63] https://apexcharts.com

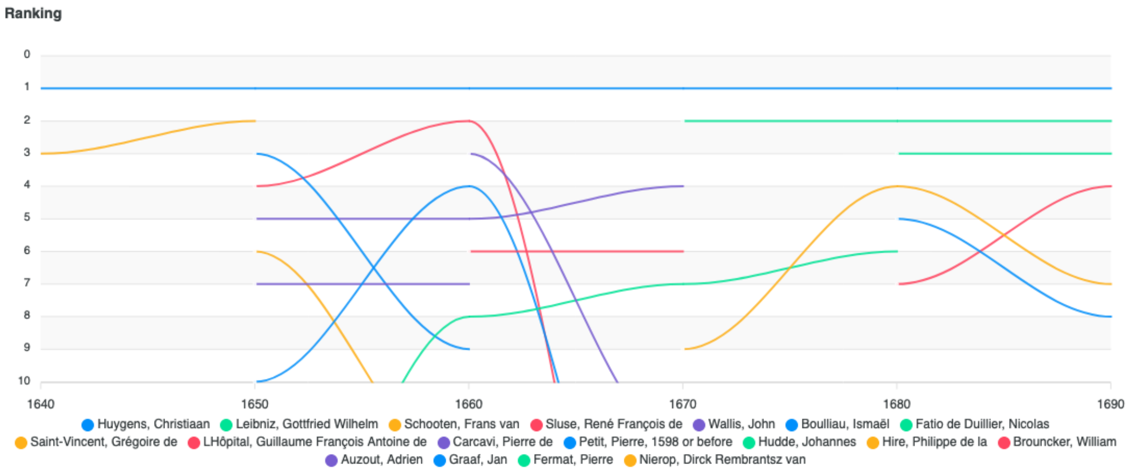


Fig. 10. The most active matematician correspondents on a timeline, as far as the data in the CKCC corpus tells

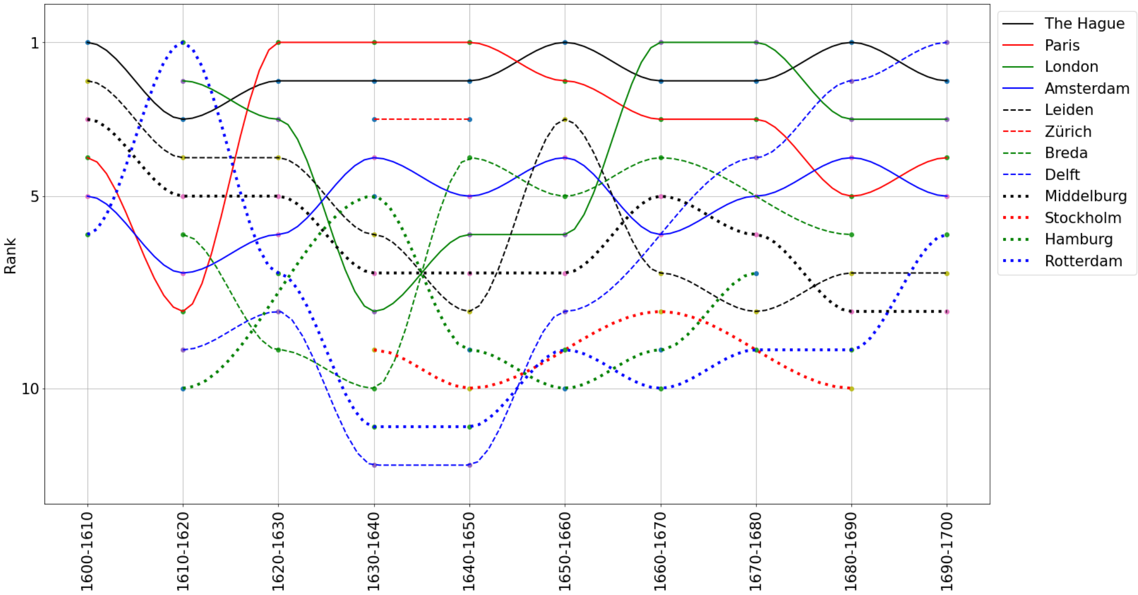


Fig. 11. The most mentioned places on a timeline

It should be noted that although analyses and visualizations are now arguably easier to create using Sampos, the real challenges of interpreting the results from a humanities point of view are still there, and close reading of the data is necessary, too. In general, lots of data literacy is needed in DH research to understand, e.g., the biases, assumptions, omissions, uncertainties, and errors in the data [32, 36]. As noted above, a particular challenge faced when analyzing epistolary data is incompleteness of the underlying collections. For example, determining the most active mathematicians in Fig. 10 is of course based on only the data that is available. There may be more active mathematicians around whose letters are not included in the dataset for one reason or another. It is also possible that the letter data for one person may not be as complete as for another, which skews the results. These problems can be mitigated when studying the collections quantitatively from a historiographical point of view, like in [42, 51]. For example, Fig. 11 presents a historiographical fact about the places mentioned in the metadata (as far as the data is correct). However, the question of what the analysis actually tells about the development of the Republic of Letter in the Netherlands in time needs much more research by a domain expert.

The example analyses presented in this paper were created as illustrations of the new technical opportunities available when using the LetterSampo framework. The idea is that this kind of analyses can point out potentially interesting patterns of data and phenomena in the underlying real world to be investigated in more detail by close reading.

6 DISCUSSION

This section presents a summary of the contributions of the paper, followed by discussions on related works, on evaluating the framework and demonstrator, and on deployment of the framework in use on the Web.

Contributions

This paper presented a model of using LOD as the basis for aggregating, harmonising, publishing, and using epistolary data for DH research. To study and test the idea, the Sampo model principles were applied to create the LetterSampo framework. It provides guidelines and software support 1) to create LOD services and intelligent semantic portals for epistolary data, 2) to facilitate data analysis of semantic Big Data, 3) to serve back the aggregated data to their original creators for DH research, and to 4) foster application development of third parties by using a LOD service with its SPARQL API and other services. As proof-of-concept, the framework was implemented using Sampo-UI and applied to three corpora, CKCC, correspSearch, and EMLO. CKCC and correspSearch LOD were published on the Web using open licenses as SPARQL endpoints with additional services at the Linked Data Finland platform, and as data dumps in the Zenodo service.

Also a portal demonstrator based on the aggregated CKCC and correspSearch LOD was published on the Web for public use. By using the LetterSampo framework new portal instances can be created easily if the data model used conforms to that of the LetterSampo framework. This paper focused on presenting, discussing, and illustrating design principles for publishing and using epistolary data as linked data on the Web, not on presenting analysis results of particular datasets.

Related Work

There are several systems and services around for storing, aggregating, searching, and studying epistolary data on the Web, as introduced in Section 1, but with the focus on serving information for humans to read, not for machines to support DH research and application development. However, in some cases, such as the web service of correspSearch, 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 [45] in relation to EMLO. The Norwegian Correspondences project [41] as well as the Constellations of Correspondence project [44] aim to base their web services on linked data.

The principles behind the Sampo model and LetterSampo framework have been explored and developed before in different contexts. For example, the notion of collaborative content creation by data linking is a fundamental idea behind the Linked Open Data Cloud movement[64] and has been developed also in various other settings, e.g., in ResearchSpace[65]. The two step filter-analyze usage model used in Sampo portals is also used, e.g., in prosopographical research [50] (without the faceted search component). Providing multiple analyses and visualizations to a set of filtered search results has been used in other portals, such as the ePistolarium[66] [39] for epistolary data, and using multiple perspectives to data has been studied as an approach in decision making [31]. Faceted search [43, 46], also known as ”view-based search” and ”dynamic ontologies”, is a well-known paradigm for explorative search and browsing [33] in computer science and information retrieval, based on S. R. Ranaganathan’s original ideas of faceted classification in Library Science. The novelty of the Sampo model and LetterSampo framework lies in consolidating several ideas together and in operationalizing them for developing applications in Digital Humanities, in this case for epistolary research. This is something that the field of the Semantic Web is arguably missing [11]. Based on Linked Data principles, the LetterSampo framework is compatible with the FAIR principles for creating Findable, Accessible, Interoperable, and Re-usable data[67].

Evaluation

The goal of this paper has been to illustrate by examples the opportunities and challenges of the LetterSampo framework. The data and demonstrator are now openly ready to be tested as a tool for research in epistolary studies. At the same time, the usability of the LetterSampo demonstrator UI could be evaluated by external users. Burrows et al. [2] report on evaluating a Sampo-UI-based UI using the Mapping Manuscript Migrations (MMM) portal [23] where pre-modern manuscript collections are used as data. The results suggested promising usability of the Sampo model and its UI logic from an end user’s point of view, but of course the MMM portal is different from the LetterSampo demonstrator. An empirical indication of the usability of the Sampo systems is that they have now been used in several different online CH portals that have had over million users on the Web in total[68].

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 [5, 8]. 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.

As for computational complexity, the Sampo-UI tool has been shown to scale up to hundreds of thousands of search objects (class instances) but the complexity depends on the data model used and how many and how large hierarchical facets are used [27]. A computationally demanding task in faceted search is pre-computing the hit counts for each facet category after each filtering step. When dealing with very large instance sets it is possible to force the user to constrain the search space first and use faceted search only after that. In NameSampo, for example, over two million placenames related to places are considered, and the search focus is initially constrained by limiting the area in question on a map or by text search on names [28].

Deploying the Framework

The LetterSampo framework uses linked data that has typically been aggregated from heterogenous, distributed data sources. From a data production point of view, lots of work by the portal developers is then needed in harmonizing, aligning, and linking the local datasets unambiguously. Arguably a better way to go would be to first agree upon using a

[64] https://lod-cloud.net
[65] https://www.researchspace.org
[66] http://ckcc.huygens.knaw.nl
[67] https://www.go-fair.org/fair-principles/
[68] Information about the Sampo portal series is available at: https://seco.cs.aalto.fi/applications/sampo/.

global data infrastructure, i.e., the data models and domain ontologies used, and distribute the linked data creation to the participating local data providers, where the best knowledge about their data is available [15] (cf. Fig. 1).

In our case studies, the data aggregation work had already been done by the corpus aggregators using conventional methods. However, we believe that linked data would be useful for this task, too. Distributed content creation of linked data can be supported by ontology services and other tools on top of the shared infrastructure as suggested in [20]. A challenge here is that legacy systems in use do not usually support Linked Data, and the technology is still fairly new and not consistently established in IT departments. The most important challenge is, however, that using the new model requires greater collaboration and mutual agreements between the participating organizations, which complicates the process. One has to take into consideration the shared data models, ontologies, and vocabularies used by the community, not only one's own preferred standards and practices. However, since in this case the final goal of the community is to create a global view to historical epistolary data, such as the RofL, it is a better idea to prevent interoperability problems to arise by a Linked Data infrastructure than to try to solve them afterwards when the damage is already done [15]. As Albert Einstein put it: *Intellectuals solve problems, geniuses prevent them*.

ACKNOWLEDGMENTS

Esko Ikkala and Heikki Rantala contributed in developing and applying the Sampo model and Sampo-UI framework in various Sampo portals. Thanks to Charles van den Heuvel and Dirk van Miert for discussions and providing the CKCC data, and to Stefan Dumont for providing the metadata of the correspSearch system. We also thank Arno Bosse, Howard Hotson, Miranda Lewis, and Eetu Mäkelä for collaborations related to the EMLO data during the *Cultures of Knowledge* project at the University of Oxford, funded by the Mellon Foundation, as well as other colleagues in the EU COST Action project *Reassembling the Republic of Letters*[69]. Collaborations with Javier Ureña-Carrion and Mikko Kivelä are acknowledged regarding network analysis of RofL data. The work of this paper is partly funded by the EU project InTaVia: In/Tangible European Heritage[70]. Our work is also based on results of the *Open Science and Research Programme*[71], funded by the Ministry of Education and Culture of Finland. CSC – IT Center for Science, Finland has provided computational resources for the work.

REFERENCES

[1] Toby Burrows, Laura Cleaver, Doug Emery, Eero Hyvönen, Mikko Koho, Lynn Ransom, Emma Thomson, and Hanno Wijsman. 2022. Medieval manuscripts and their migrations: Using SPARQL to investigate the research potential of an aggregated Knowledge Graph. *Digital Medievalist* (2022). https://doi.org/10.16995/dm.8064 Prepint: https://seco.cs.aalto.fi/publications/2021/burrows-et-al-digital-medievalist-2021.pdf.

[2] Toby Burrows, Nicole Bergk Pinto, Mahaut Cazals, Alexandre Gaudin, and Hanno Wijsman. 2020. Evaluating a Semantic Portal for the "Mapping Manuscript Migrations" Project. *DigItalia* 2 (2020), 178–185. http://digitalia.sbn.it/article/view/2643

[3] Stefan Dumont. 2016. correspSearch – Connecting Scholarly Editions of Letters. *Journal of the Text Encoding Initiative* 10 (2016). https://doi.org/10.4000/jtei.1742

[4] Stefan Dumont, Sascha Grabsch, and Jonas Müller-Laackman. 2021. correspSearch – Connecting Scholarly Editions of Correspondence (2.0.0) [Web service]. Berlin–Brandenburg Academy of Sciences and Humanities. https://correspSearch.net

[5] J. English, M. Hearst, R. Sinha, K. Swearingen, and K.-P. Lee. 2003. *Flexible search and navigation using faceted metadata*. Technical Report. University of Berkeley, School of Information Management and Systems.

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

[7] Aneeq Hashmi, Faraz Zaidi, Arnaud Sallaberry, and Tariq Mehmood. 2012. Are all social networks structurally similar?. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, IEEE, 310–314. https://doi.org/10.1109/asonam.2012.59

[69] http://www.republicofletters.net
[70] https://intavia.eu
[71] http://openscience.fi

[8] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee. 2002. Finding the flow in web site search. *CACM* 45, 9 (2002), 42–49.

[9] Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Morgan & Claypool, Palo Alto, CA. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

[10] Sara Heydari, Sam G. Roberts, Robin I. M. Dunbar, and Jari Saramäki. 2018. Multichannel social signatures and persistent features of ego networks. *Applied Network Science* 3, 1 (May 2018). https://doi.org/10.1007/s41109-018-0065-4

[11] Pascal Hitzler. 2021. A Review of the Semantic Web Field. *Commun. ACM* 64, 2 (Jan. 2021), 76–83. https://doi.org/10.1145/3397512

[12] Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. 2010. *Foundations of Semantic Web technologies*. Springer–Verlag.

[13] Howard Hotson and Miranda Lewis. 2022. *Early Modern Letters Online*. University of Oxford. http://emlo.bodleian.ox.ac.uk/home

[14] Howard Hotson and Thomas Wallnig (Eds.). 2019. *Reassembling the Republic of Letters in the Digital Age*. Göttingen University Press. https://doi.org/10.17875/gup2019-1146

[15] Eero Hyvönen. 2010. Preventing Interoperability Problems Instead of Solving Them. *Semantic Web – Interoperability, Usability, Applicability* 1, 1–2 (2010), 33–37. https://doi.org/10.3233/SW-2010-0014

[16] Eero Hyvönen. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Morgan & Claypool, Palo Alto, CA, USA. https://doi.org/10.2200/S00452ED1V01Y201210WBE003

[17] Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. 2019. BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research. In *The Semantic Web: ESWC 2019*. Springer–Verlag. https://doi.org/10.1007/978-3-030-21348-0f_37

[18] E. Hyvönen, S. Saarela, and K. Viljanen. 2004. Application of ontology based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10–12, Heraklion, Greece*. Springer–Verlag, 92–106. https://seco.cs.aalto.fi/publications/2004/hyvonen-saarela-et-al-application-of-ontology-techniques-2004.pdf

[19] Eero Hyvönen, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä. 2014. 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, 226–230. https://doi.org/10.1007/978-3-319-11955-7_24

[20] E. Hyvönen, K. Viljanen, J. Tuominen, and K. Seppälä. 2008. Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*. Springer–Verlag, 95–109. https://doi.org/10.1007/978-3-540-68234-9_10

[21] Eero Hyvönen. 2020. Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web – Interoperability, Usability, Applicability* 11, 1 (2020), 187–193. https://doi.org/10.3233/SW-190386

[22] Eero Hyvönen. 2022. Digital Humanities on the Semantic Web: Sampo Model and Portal Series. *Semantic Web – Interoperability, Usability, Applicability* (2022). http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series Accepted.

[23] Eero Hyvönen, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 2021. Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research. In *Semantic Web. Proceedings of the The 20th International Semantic Web Conference (ISWC 2021)*. Springer–Verlag. https://doi.org/10.1007/978-3-030-88361-4_36

[24] Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. 2022. *LetterSampo CKCC data publication*. Zenodo. https://doi.org/10.5281/zenodo.6631385

[25] Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. 2022. *LetterSampo correspSearch data publication*. Zenodo. https://doi.org/10.5281/zenodo.6631357

[26] Eero Hyvönen, Heikki Rantala, Esko Ikkala, Mikko Koho, Jouni Tuominen, Babatunde Anafi, Suzie Thomas, Anna Wessman, Eljas Oksanen, Ville Rohiola, Jutta Kuitunen, and Minna Ryyppö. 2021. Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework. *Antiquity, A Review of World Archaeology* 95, 382 (July 2021). https://doi.org/10.15184/aqy.2021.87

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

[28] Esko Ikkala, Jouni Tuominen, Jaakko Raunamaa, Tiina Aalto, Terhi Ainiala, Helinä Uusitalo, and Eero Hyvönen. 2018. NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research. In *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities* (Seattle, WA, USA) *(GeoHumanities'18)*. ACM, New York, NY, USA, Article 2, 9 pages. https://doi.org/10.1145/3282933.3282936

[29] Mikko Koho, Toby Burrows, Eero Hyvönen, Esko Ikkala, Kevin Page, Lynn Ransom, Jouni Tuominen, Doug Emery, Mitch Fraas, Benjamin Heller, David Lewis, Andrew Morrison, Guillaume Porte, Emma Thomson, Athanasios Velios, and Hanno Wijsman. 2022. Harmonizing and Publishing Heterogeneous Pre-Modern Manuscript Metadata as Linked Open Data. *Journal of the Association for Information Science and Technology (JASIST)* 73, 2 (2022), 240–257. https://doi.org/10.1002/asi.24499

[30] Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2021. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. *Semantic Web* 12, 2 (Jan 2021), 265–278. https://doi.org/10.3233/SW-200392

[31] Harold A. Linstone. 1989. Multiple perspectives: Concept, applications, and user guidelines. *Systems practice* 2, 3 (1989), 307–331. https://doi.org/10.1007/BF01059977

[32] Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen. 2020. Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*. CEUR Workshop Proceedings, 81–96. http://ceur-ws.org/Vol-2612/paper6.pdf

[33] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. of ACM* 49, 4 (April 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[34] Willard McCarty. 2005. *Humanities Computing*. Palgrave, London.

[35] Albert Meroño-Peñuela and Rinke Hoekstra. 2016. grlc makes GitHub taste like linked data APIs. In *The Semantic Web. ESWC 2016*. Springer–Verlag, 342–353. https://doi.org/10.1007/978-3-319-47602-5_48

[36] Franco Moretti. 2013. *Distant reading*. Verso Books.

[37] Eljas Oksanen, Heikki Rantala, Jouni Tuominen, Michael Lewis, David Wigg-Wolf, Frida Ehrnsten, and Eero Hyvönen. 2022. Digital Humanities Solutions for Pan-European Numismatic and Archaeological Heritage Based on Linked Open Data. In *6th Digital Humanities in Nordic and Baltic Countries Conference*. CEUR Workshop Proceedings. Forth-coming, preprint: https://seco.cs.aalto.fi/publications/2022/oksanen-et-al-diginuma-dhnb-2022.pdf.

[38] Evelien Otte and Ronald Rousseau. 2002. Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science* 28, 6 (2002), 441–453.

[39] Walter Ravenek, Charles van den Heuvel, and Guido Gerritsen. 2017. The ePistolarium: Origins and Techniques. In *CLARIN in the Low Countries*, Arjan van Hessen and Jan Odijk (Eds.). Ubiquity Press, 317–323. https://doi.org/10.5334/bbi

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

[41] Annika Rockenberger, Ellen Nessheim Wiger, Mette Refslund Witting, Hilde Bøe, Evelyn Irene Thor, Ove Joralf Wolden, Marianne Paasche, Ola Søndenå, and Philipp Conzett. 2019. Norwegian Correspondences and Linked Open Data. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (CEUR Workshop Proceedings, Vol. 2364)*, Costanza Navarretta, Manex Agirrezabal, and Bente Maegaard (Eds.). 365–375. http://ceur-ws.org/Vol-2364/33_paper.pdf

[42] Minna Tamper, Petri Leskinen, Eero Hyvönen, Risto Valjus, and Kirsi Keravuori. 2021. Analyzing Biography Collection Historiographically as Linked Data: Case National Biography of Finland. *Semantic Web – Interoperability, Usability, Applicability* (2021). Forth-coming, preprint: https://seco.cs.aalto.fi/publications/2021/tamper-et-al-bs-2021.pdf.

[43] D. Tunkelang. 2009. Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1, 1 (2009), 1–80.

[44] Jouni Tuominen, 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*. CEUR Workshop Proceedings. Forth-coming, preprint: https://seco.cs.aalto.fi/publications/2022/tuominen-et-al-coco-dhnb-2022.pdf.

[45] Jouni Tuominen, Eetu Mäkelä, Eero Hyvönen, Arno Bosse, Miranda Lewis, and Howard Hotson. 2018. Reassembling the Republic of Letters - A Linked Data Approach. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)* (Helsinki, Finland). CEUR Workshop Proceedings, vol. 2084, 76–88. http://www.ceur-ws.org/Vol-2084/paper6.pdf

[46] Yannis Tzitzikas, Nikos Manolis, and Panagiotis Papadakos. 2017. Faceted exploration of RDF/S datasets: a survey. *Journal of Intelligent Information Systems* 48, 2 (2017), 329–364. https://doi.org/10.1007/s10844-016-0413-8

[47] Javier Ureña-Carrion, 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. *Applied Network Science* (2022). https://arxiv.org/abs/2112.04336v1 In press.

[48] Charles van den Heuvel. 2015. Mapping Knowledge Exchange in Early Modern Europe: Intellectual and Technological Geographies and Network Representations. *International Journal of Humanities and Arts Computing* 9, 1 (3 2015), 95–114. https://doi.org/10.3366/ijhac.2015.0140

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

[50] Koenraad Verboven, Myriam Carlier, and Jan Dumolyn. 2007. A short manual to the art of prosopography. In *Prosopography approaches and applications. A handbook*. Unit for Prosopographical Research (Linacre College), 35–70. https://doi.org/1854/8212

[51] Christopher N. Warren. 2018. Historiography's Two Voices: Data Infrastructure and History at Scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics* 1, 2 (2018), 1–31. https://doi.org/10.22148/16.028

[52] Marcia Zeng and Jian Qin. 2022. *Metadata, Third Edition*. ALA Neal-Schuman, Chicago.