Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web

Petri Leskinen$^{1\,[0000-0003-2327-6942]}$, Javier Ureña-Carrion$^{3\,[0000-0003-2763-9747]}$, Jouni Tuominen$^{1,2\,[0000-0003-4789-5676]}$, Mikko Kivelä$^{3\,[0000-0003-2049-1954]}$, and Eero Hyvönen$^{1,2\,[0000-0003-1695-5840]}$

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi
[3] Complex Systems Group, Aalto University, Finland

Abstract. Communication data between people is a rich source for insights into societies and organizations in areas ranging from research on history to investigations on fraudulent behavior. These data are typically heterogeneous datasets where communication networks between people and the times and geographical locations they take place are important aspects. We argue that these features make the area of temporal communications a promising application case for Linked Data (LD) -based methods combined with temporal network analyses. A key result of this paper is to show how to create and publish a global Linked Open dataset and data service about historical epistolary data, based on distributed data from several international heterogeneous data sources, that can be enriched by data linking and reasoning, and that can be served back for the research community as an open infrastructure, a data service, and a semantic portal for further study in Digital Humanities (DH). A framework for this purpose is presented for publishing and analyzing communication network data, based on recent advances in analysis of temporal communication networks and the behavioral patterns commonly found in them. We applied the framework to creating and publishing two open LD services (CC BY 4.0): 1) the data schema, dataset, and data service of the Dutch CKCC corpus (of ca. 20 000 letters) and 2) the pan-European correspSearch corpus (of ca. 135 000 letters) related to the Republic of Letters (1500–1800). To evaluate and demonstrate the usability of the new data services in DH research, a semantic portal was implemented on top of their SPARQL endpoint demonstrating re-usability, flexibility, and feasibility of the Linked Data approach from a DH research perspective.

Keywords: Semantic Web, Linked Open Data, Digital Humanities, Network Science, Early Modern, Correspondence

1 Introduction

Since the revolution in network science around 20 years ago [35,36,30], this field of research has been extremely successful explaining various phenomena and fundamental concepts in a wide array of systems from societies to brain and cellular biology.

The tools and ideas developed for network analysis range from ideas characterising the whole network to diagnostics computed for individual nodes in the network, such as centrality measures, node roles, and local clustering coefficients. However, these tools are often mainly used by the network scientists as they are difficult to use for the domain experts: accessing them requires programming skills or at least specialised software that relates the often heterogeneous network data and metadata to the questions that are important for the domain experts. On the other hand, there is a need to make the rich datasets created by historians in Digital Humanities (DH) and the Linked Data community available for the network scientists.

This paper builds on the idea that Semantic Web technologies[4] [10] and linked data [7,14] can be a solution to these problems. The graph-based RDF data model underlying the Semantic Web is a perfect match for representing network data, and Linked Data publishing [7] can be used for making the data available for researchers in humanities with some skills on using SPARQL[5] queries or on programming with SPARQL endpoints. Furthermore, ready to use portal solutions for data analysis can be implemented for DH based on such data services [16]. The idea is that by combining the flexibility of publishing and using LD with the tools of network science can help domain experts to tackle massive network data in fruitful manner with little or no expertise in programming. Furthermore, the LD created can be served back to the research community for further research and application development in a disciplined and well-defined way by using the Semantic Web methodology [11] the practical LD publishing principles including SPARQL endpoints.

Table 1. Datasets analyzed and compared in this paper

Dataset	**Content**
CKCC	Epistolary data of the CKCC corpus of the Huygens Institute in the Netherlands, an aggregated collection of ca 20 000 Dutch correspondences [8,24] related to the Republic of Letters [24,13]
correspSearch	Epistolary data 1510–1991 of 135 000 letters aggregated by the correspSearch project at the Berlin Brandenburg Academy of Sciences and Humanities [2]

To test and demonstrate this approach in practise, this paper focuses on communication networks that are represented as temporal networks, a rapidly developing subfield of network science [30,12]. Two datasets of historical epistolary data listed in Table 1 are used as case study examples. Temporal networks are a specific type of networks that carry information on the activation times of the links in addition to the topological structure of the networks. In communication networks this means that we do not only consider who has been in contact with whom, but also the exact time instances at which the communication has taken place. This not only adds complications related to how the various methods and measures are generalised for temporal networks, but also creates possibilities of new types of network analysis. For example, in communication networks

[4] `https://www.w3.org/standards/semanticweb/`

[5] `https://www.w3.org/TR/sparql11-query/`

it has been found that the individuals are in contact in a bursty manner [4,20] and they distribute their communication efforts via patterns, known as social signatures, that are specific to each individual [29,9]. These phenomena are understood in terms of statistical laws found in anonymized data, but much less attention has been given on how such features translate to interpretations of individual relationships or people. Here we introduce a method for giving access to these state-of-the-art network analysis methods to domain experts, who work through the massive databases of communications using theoretically grounded analysis tools. It should be noted that this paper focuses on presenting a technical framework and approach for applying network analysis and LD technology to publishing and using historical epistolary data in research, not on particular domain specific analyses of the datasets from a humanities point of view. This remains a proposed topic of further research using the approach and tooling presented.

The paper is organized as follows. First, related work in epistolary historical network studies and temporal network analysis and systems are discussed to contextualize the work of this paper. Next a new data model and linked data sets conforming to it are presented as well as a LD service platform for publishing them, based on extending the traditional 5-star model to a 7-star model. After this, examples of network analyses using the linked data and SPARQL endpoint are presented. To test and demonstrate usability of the new data resource and data service even further, a semantic portal on top of the data service is presented with examples of data analyses. In conclusion, contributions of the paper and challenges of the proposed approach are summarized and discussed.

2 Related Work

Epistolary Historical Networks 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*, 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. Data sources of early Early Modern learned correspondences are proliferating rapidly, including, e.g., Europeana[6], Kalliope[7], The Catalogus Epistularum Neerlandicarum[8], Electronic Enlightenment[9], ePistolarium[10], SKILLNET[11], correspSearch[12], the Mapping the Republic of Letters project[13], and Early Modern Letters Online (EMLO)[14] [8,24,13]. Visualizing the correspondences has been studied in the Mapping the Republic of Letters project[15].

[6] http://www.europeana.eu

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

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

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

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

[11] https://skillnet.nl

[12] https://correspsearch.net

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

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

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

The idea of representing epistolary data as a LD service was introduced in [31] and its application to DH research is discussed in [15] pointing out the analogy between RofL and Linked Open Data movement with some tooling, data analyses, and visualizations as examples. In this paper, the idea of using the linked data service is developed and discussed further from a network analytic perspective, in relation to the correspondences of the Republic of Letters 1500–1800 [13] in the Netherlands and Germany. We demonstrate flexibility and scientific potential of using an epistolary linked data service for research in the following ways: 1) Firstly, by transforming and downloading the data into a suitable form, network analytic tools developed originally for different purposes, in our case for contemporary communication data, can be re-used, making it possible to apply them to historical epistolary networks, too. 2) Secondly, based on the Sampo model [16] and Sampo-UI framework [19], the data service can be integrated seamlessly with tooling for DH research making network analyses possible for researchers in Humanities who often lack programming experience. 3) Thirdly, it is shown how the LD data service resource can be used for solving Digital Humanities problems with little programming experience using online programming services, such as Google CoLab[16] and Jupyter[17].

Temporal Network Analysis In the past few decades, communication data has become a relevant resource to understand the underlying social networks [26,30]. In such cases, auto-recorded logs of pairwise interactions are modelled to construct a communication network, thus allowing the analysis of large-scale societal interactions and behavioural patterns. Here we focus on using epistolary Linked Data about communications to analyse historical correspondence networks of epistolary data but the methodology can equally well be used for modern communication networks, such as those from mobile phone logs, emails and social media platforms [32]. We take two main approaches to analyse such communication datasets, first a static approach — where we establish a link between two people if there have been epistolary contacts between them—, and a temporal approach, where we examine the distribution of dyadic interactions and as well as behavioural features that characterize the way that people communicate. We note, however, that while most modern datasets attempt to capture all communication within a communication channel (e.g., all emails or other communications within an organisation [3,1,37]), we cannot assume this to be true for historical data, since its collection is not automated, but implies herculean compilation efforts by researchers.

For the static approach, we build a network from dyadic interactions by filtering out data that does not fall within a certain period (e.g., 1700–1750), although this process can also be performed using geographical information. The reason for doing this is because the interpretation of most social network analysis assumes that any two nodes may interact, which is not possible in large observation periods. We create a link between two people if there has been epistolary contact, and we assign a proxy for the strength of a tie based on the total number of contacts [26,33]. We may then examine large-scale properties of the resulting networks, including the degree distribution (i.e., the number of contacts of each node), different centrality measures (i.e., metrics to cap-

[16] `https://colab.research.google.com`

[17] `https://jupyter.org`

ture the relative importance of nodes within the network), or measures of the existence of communities or other types of structures.

For our temporal approach, we analyze the distribution of time-sequences of events of dyadic interactions, its relationship to the local network structure, as well as behavioural characteristics of how individual people communicate with their neighbours. In the first case, we take a sequence of times of contacts and apply time-series analysis techniques, noting the average time between contacts, the variability of timings between contacts — known to be quite large for human communication—, as well as an attempt to capture the overall period in which communication was active. Then, we take such features and link them to the local structure. From a sociological standpoint, it has been proposed that *strong ties* tend to be buried in overlapping circles of friends, akin to small communities, which *weak ties* serve more as bridges between such communities [5]. Since it is not possible to directly observe the strength of a tie, it is possible to use different temporal features as proxies, and we thus see that stronger ties do tend to be embedded in circles of friends [33]. Last, we focus on how individuals divide their contacting behaviour across their different neighbors in ego networks. This analysis is built on the observations that people have persistent *social signatures* [29,9]—how they divide their communication is fairly consistent so that in different time periods their particular top-ranked contacts might change, but they will contact their top-ranked contacts similarly.

Using Linked Data for Network Analysis The idea of using Linked Data graphs in network science is intuitive, natural, and not new. For example, in [6] linked data is transformed for network analysis for the LinkedDataLens system. In [27] RDF data is used for Social Network Analysis. Linked data from different sources can be aggregated into larger networks and enriched by each other and by reasoning new triples, i.e., connections in the network. SPARQL queries and SPARQL CONSTRUCT can be used in flexible ways for network data transformations and creating tabular formats widely used. To facilitate network analysis and visualizations of RDF data there are tools available, such as the Semantic Web Import Plugin plugin[18] available for Gephi[19], arguably the leading visualization and exploration software for all kinds of graphs and networks. Applications of Gephi include, for example, Exploratory Data Analysis, Link Analysis, Social Network Analysis, and Biological Network analysis. A major contribution of our paper is to apply network analysis in a novel application domain for analysing historical epistolary communication networks, and especially by using temporal network analysis. For this purpose, a new LOD resource is presented and used.

3 A Linked Data Model and Service for Epistolary Data

This paper makes use of the epistolary datasets listed in Table 1. In our work these datasets were transformed into Linked Data and published according to the Linked Data publishing principles and other best practices of W3C [7], including, e.g., content negotiation and provision of a SPARQL endpoint. The CKCC corpus is to the best of our knowledge the first public linked open dataset on the Web on historical epistolary data;

[18] `https://www.w3.org/2001/sw/wiki/GephiSemanticWebImportPlugin`

[19] `https://gephi.org/`

opening the publication of the correspSearch data in a similar way is done after getting a confirmation of the open license from the data owner.

Data Model for Linked Epistolary Data By transforming the epistolary data into RDF we aimed to create knowledge graphs that include not only communication networks but also prosopographical data about the people and organizations involved. For this purpose a customised RDF-based metadata schema was created. The schema contains four different, interlinked classes: Letter, Actor, Tie, and Place as described in Table 2. Here the default namespace is our own (*ckcc-schema*), *rdfs* refers to the RDF Schema[20], *crm* to the CIDOC CRM Schema[21], and *xsd* to the XML Schema of W3C[22].

In the epistolary dataset, instances of the class Actor can be either people or groups. The sent letters by each actor are announced using the property :created. Each letter is modeled as an instance of the class Letter that has seven properties describing the letter. A letter is linked with its recipients using the property :was_addressed_to, to places of sending and receiving using the properties :was_sent_from and :was_sent_to, and to related timespan with crm:P4_has_time-span. Furthermore, a letter instance is enriched with information about the data source and a human-readable description. The correspondences between two actors are modeled as instances of the class Tie. Each of the Tie instances is linked to the two actors and likewise each letter is linked to the corresponding tie. Using the Tie instance simplifies the database queries e.g. in cases of querying all the letters between the two actors. In addition, this model facilitates to adding precalculated network metrics such as node degrees and centrality measures to the data model. In addition, the data set also contains precalculated values for the time of flourishing for each actor, e.g. the time period when the actor has been active in letters correspondences. The resources in the domain ontology of the places consist of place labels, the coordinate information, and the hierarchy built with the property skos:broader. Finally, the timespans follow the four point model, e.g. with xsd:dateTime values indicating the earliest and latest moments for the beginning and the end.

The two datasets, CKCC and correspSearch, were converted from different source formats. CKCC is an extract from an existing RDF dataset, while the correspSearch data was converted from a source in JSON format. In these datasets both the actor and place resources had linkage to other LOD cloud databases, e.g., Wikidata, VIAF, Early Modern Letters Online project (EMLO), or database of Deutsche Nationalbibliothek[23] (GND). This existing linkage was used for two main purposes. First, in the current data publication, the resources in the datasets where reconciled based on the links, e.g. the actors or places refer to the same entity, if they point to the same external link. Secondly, the external databases was used to enrich our data e.g. with images of actors and coordinates of the places. In our work, the "FAIR guiding principles for scientific data management and stewardship" of publishing Findable, Accessible, Interoperable, and Re-usable data are used[24].

[20] https://www.w3.org/TR/rdf-schema/

[21] http://www.cidoc-crm.org

[22] https://www.w3.org/XML/Schema

[23] https://www.dnb.de/EN/Home/home_node.html

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

Table 2. RDF schema for Letter, Actor, Tie, and Place. Column *C* marks the cardinality of the element. Fields inferred from the data are marked with *cursive* text.

Element URL	C	Range	Meaning of the value
ACTOR			
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
LETTER			
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
crm:P4_has_time-span	0..1	crm:E52_Time-Span	Time of sending
:source	1..n	rdfs:Resource	Used data source
:in_tie	1	:Tie	*Correspondence in which this letter belongs to*
TIE			
:actor1	1	crm:E39_Actor	First correspondent
:actor2	1	crm:E39_Actor	Second correspondent
:num_letters	1	xsd:integer	*Number of letters in this correspondence*
skos:prefLabel	1	xsd:string	Preferable label
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
TIMESPAN			
crm:P82a_begin_of_the_begin	0..1	xsd:dateTime	Earliest time for the beginning
crm:P81a_end_of_the_begin	0..1	xsd:dateTime	Latest time for the beginning
crm:P81b_begin_of_the_end	0..1	xsd:dateTime	Earliest time for the end
crm:P82b_end_of_the_end	0..1	xsd:dateTime	Latest time for the end
skos:prefLabel	1	xsd:string	Preferable label

Using the Linked Data and Data Service The data can be used for research via 1) ready-to-use tools available on a semantic portal or 2) by using the underlying SPARQL endpoint with external tools, based on a framework called *LetterSampo* [15]. The SPARQL endpoint can be used directly in Digital Humanities research using, e.g., YASGUI[25] [28] and Python scripting in Google Colab and Jupyter notebooks. The endpoint can also be used for filtering and downloading the data in different forms, such as in tabular CSV format, for external data-analysis tools, in our case for network analyses.

This framework is used for creating data services and semantic portals[26] based on the Sampo model [16] for sharing collaboratively enriched linked open data using a shared ontology infrastructure. The portals host ready-to-use data-analytic tools for

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

[26] `https://seco.cs.aalto.fi/applications/sampo/`

Digital Humanities research, as suggested in [17]. The Sampo-UI framework [19] is used as the interface model and as the full stack JavaScript tool was used for implementing the user interface. Sampo portals are based—from a data perspective—solely on querying the SPARQL endpoint from the client side using JavaScript. They demonstrate the idea that versatile web applications can be implemented by separating the application logic and data services via SPARQL API, which arguably facilitates developing new applications efficiently by re-using the same data.

New Resources on the Web The CKCC knowledge graph[27] as well as the correspSearch knowledge graph[28] have been published on the Linked Data Finland platform LDF.fi [18]. Both dataset are also available at Zenodo[29] LDF.fi uses the "7-star" model for LD deployment [18] that extends the "5-star" model[30] coined by Tim Berners-Lee: to enhance re-usability of LD, the sixth star is given if the data is published with its schema and the 7th star if validation results of the data using the schema are provided. LDF.fi is powered by the Fuseki SPARQL server[31] and Varnish Cache web application accelerator[32] for routing URIs, content negotiation, and caching. The portal user interface was implemented by the Sampo-UI framework [19]. The system uses Docker microservice architecture containers[33]. By using containers, the services can be migrated to another computing environment easily, and third parties can re-use and run the services on their own. The architecture also allows for horizontal scaling for high availability, by starting new container replicas on demand.

4 Network Analyses Using the Linked Data Service

In this section we first show some general network analyses results of the epistolary datasets of Table 1. After this, it is shown how the SPARQL endpoint can be used for research using querying and by programming. For these purposes, examples using the data with custom network analytic tools, Yasgui and Google Colab are presented, respectively. Finally, analyzing the data with ready-to-use tools and the two-step analysis model of the LetterSampo portal is discussed with examples.

Exporting Data for Data Analyses An simple way of reusing the data resources is to download and transfer them for the analysis tool of choice. For this purpose either data dumps from Zenodo or the SPARQL end point can be used. A benefit of using the endpoint is that the data can be filtered and even transformed dúring the download to fit better for the aimed purpose. An example of using the data resource in external network analytic tools is presented in [34]. In this case study, the linked data of CKCC

[27] The data, schema, and service are openly available (CC BY 4.40) at the homepage `https://www.ldf.fi/dataset/ckcc`.

[28] On December 9, opening the correspSearch data service was granted by the data owners and the service will appear at `https://www.ldf.fi/dataset/correspsearch`.

[29] CKCC: `https://zenodo.org/record/5970105`
correspSearch: `https://zenodo.org/record/5972316`

[30] `https://5stardata.info/en/`

[31] `https://jena.apache.org/documentation/fuseki2/`

[32] `https://varnish-cache.org`

[33] `https://www.docker.com`

and correspSearch were analysed in terms of network metrics and compared with four modern datasets of mobile phone call networks, emails, community boards, and wall-postings on a social media platform. It turned out that contemporary and historical epistolary communication networks resemblance each other strikingly even if the media were quite different.

General Analyses on Epistolary Networks The network portal also shows the precalculated centrality measures for each actor. First, a correspondence network was created from the RDF data and thereafter the measures where calculated using the Python library NetworkX[34]. These measures take both the CKCC and the correspSearch datasets into consideration.

Table 3. Precalculated network measures for René Descartes

Measure	Value	Rank
Betweenness Centrality	0.00930	6
Clique Number	4	14
Clustering Coefficient	0.000162	380
Core Number	7	1
Eigenvector Centrality	0.064	5
Number of Correspondences	92	12
Pagerank Centrality	0.00417	23
Weighted In-Degree	164	16
Weighted Out-Degree	585	5

An example of the measures for Descartes are listed in Table 3. In the table, e.g., the Clique Number with a value of 4 indicates that Descartes is a part of complete subgraph where all the nodes have a degree of 4, and the rank of 14 indicates that there are 13 larger cliques in the entire network. The Weighted Out- and In-Degrees correspond to the total number of sent and received letters while the number of correspondences equals the unweighted node degree. Also the actor perspective facet page has a sociocentric network visualization where the actors can be filtered e.g. by their gender, years of living, or data sources.

Querying the SPARQL Endpoint For the analyses presented in this article, there are basically two practices for using a SPARQL endpoint. Firstly, for showing the data results on the web portal, the tabular results of a relatively simple query are shown on the portal page. An example of such of query is shown in Fig. 1. It queries all letters sent by Descartes and shows their recipients, labels, and dates sorted by the date. Secondly, analyzing or visualizing network structures may require several database queries e.g. for separated lists of actors (nodes) and letters (edges). The actual results are thereafter calculated based on the data of these simple, straight-forward queries with spreadsheet-like results.

Using the Endpoint by Programming Due the to performance issues when attempting to render a larger network of more than, e.g., 1000 nodes in a browser, vi-

[34] `https://networkx.org`

```
PREFIX ckccs: <http://ldf.fi/schema/ckcc/>
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?source_label ?target_label ?letter_label ?date
WHERE {
  VALUES ?source { <http://ldf.fi/ckcc/actors/p300075> }

  ?source ckccs:created ?letter ;
          skos:prefLabel ?source_label .

  ?letter a ckccs:Letter ;
          ckccs:was_addressed_to ?target ;
          skos:prefLabel ?letter_label ;
          crm:P4_has_time-span/skos:prefLabel ?date .

  ?target skos:prefLabel ?target_label .

} ORDER BY ?date
```

Fig. 1. SPARQL example for querying the letters by René Descartes

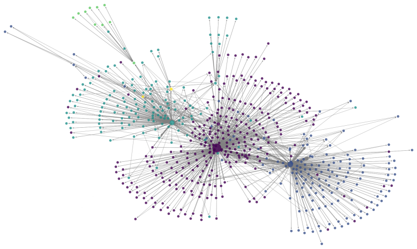


Fig. 2. Network of CKCC data

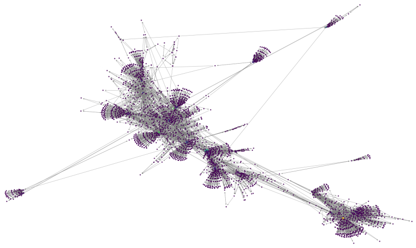


Fig. 3. Network of correspSearch data

sualizations were also performed using Python in Google Colabs environment. As an example, the largest connected component of the CKCC data is visualized in Fig. 2. The network is built around three center actors: Constantijn Huygens, Hugo de Groot, and Christiaan Huygens, who have high node degree values. On the other hand, there is a multitude of actors with low node degree. As a comparison, the correspSearch data has much more of these hubs as it is depicted in Fig. 3.

Figure 4 depicts the correspondences of Descartes on a timeline. The entire timeline is shown on the lower part of the chart. On the upper part of the chart there are separately the ten most active correspondences of Descartes and the lowest line depicts the correspondences with all the other actors. The visualization also reveals biases caused by missing information in the source data. For example, when studying the correspondence with philosopher and mathematician Marin Mersenne, it can be observed that the source collections contains 134 letters from Descartes to Mersenne, but only of five by Mersenne to Descartes.

Figure 5 depicts the most active scientists by the decades 1620–1690. The ranking is based on the total amount of sent and received letters and the data is visualized so that

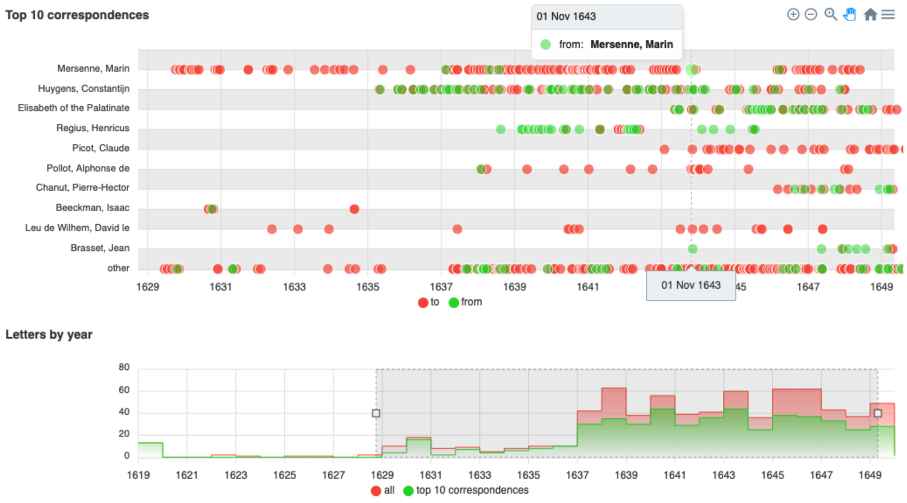


Fig. 4. Timeline depicting top 10 letter correspondences of René Descartes

the first ranking scientist is on the top of the chart. The figure depicts that from 1620 to 1640 Descartes is on the first rank, but later replaced by Christiaan Huygens.

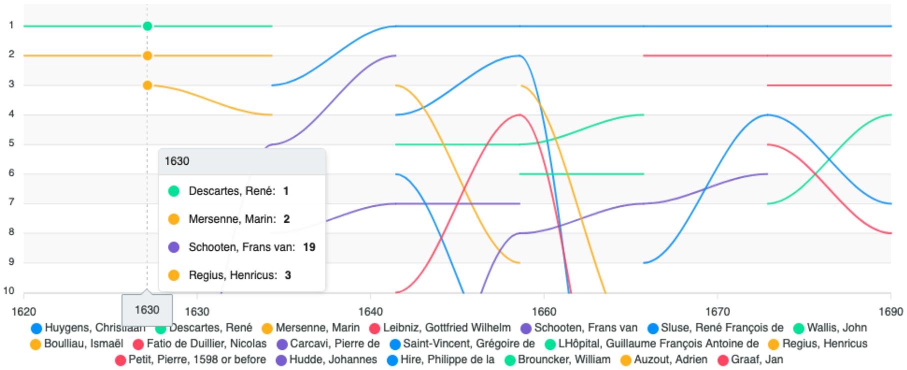
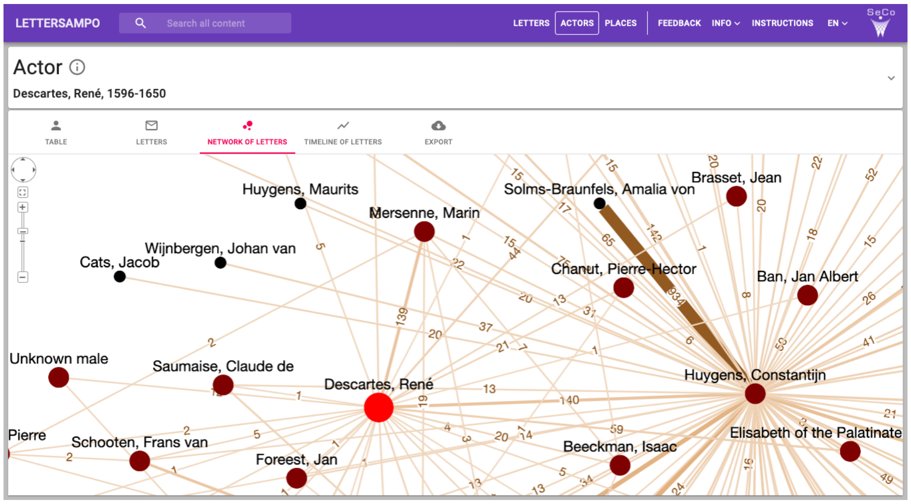


Fig. 5. Top scientists in the CKCC data during 1620–1690

Using the LetterSampo Portal The network portal provides components for visualizing the epistolary data using line charts, maps, and networks. Figure 6 depicts an egocentric network around the philosopher and scientist René Descartes. In this visualization the widths of the edges are proportional to the number of letters between the two actors while the sizes of the nodes are based on the network distance so that the main actor appears with the largest node and the most distant actors have the smallest nodes. In spite that Descartes is the center actor the poet and composer Constantijn Huygens has a higher node degree, due to the fact that the CKCC dataset contains a larger amount of letters by him.



Actor

Descartes, René, 1596-1650

Fig. 6. Letter correspondence network around René Descartes

Figure 7 depicts a visualization of the social signatures [29,9] of Descartes. Social signatures represent how individuals communicate with their neighbours in a given time. This visualization has curves for his entire time of flourishing (blue line) and separated curves during his career, e.g. the red line for time period 1643–1650. For an interval (e.g., 1631–1637, 1637–1643), we obtain a social signature by (i) computing the fraction of outgoing contacts per alter, and (ii) ranking the alters. This approach allows characterizing the relative importance of different alters in an ego network. When comparing different individuals, we find that their social signatures tend to be stable [29,9,32].

Comparing Contemporary and Historical Communication Networks As a use-case scenario, we performed a comparative analysis of historical and contemporary communication networks, with results formally introduced in [32]. In this study, the goal was to compare aspects of temporal communication networks at different granularity levels, including snapshots of static graphs, the time series of dyadic interactions, as well as ego networks. We compare these features with different contemporary communication channels, including emails, social media platforms, forums and mobile phone calls.

In brief, our goal was to analyze to what extent different behavioural features of contemporary communication networks can be found in historical datasets. We find similarities at different degrees of success. Particularly, we find evidence for the persistence

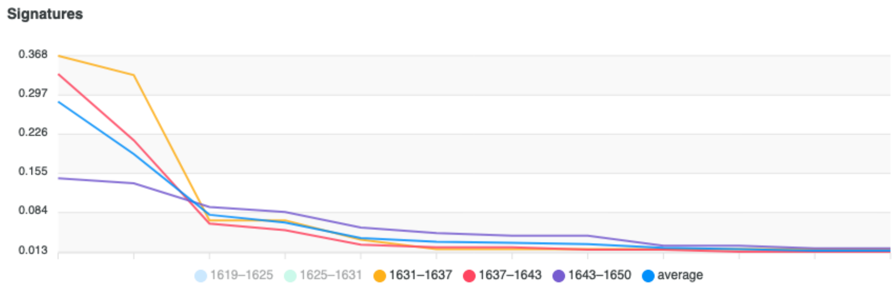


Fig. 7. Timeline depicting the social signatures of René Descartes

of social signatures in historical context, as well as the Granovetter effect for different proxies of tie strength, and important similarities in the distribution of dyadic timings. We found, however, difficulties in drawing conclusions from global network analyses, particularly given that some individuals are over-represented in historical datasets.

Regarding social signatures, our results suggest that individuals divide their communication similarly across top-ranked alters; in other words, that the social signatures of a given individual are more similar among different periods than to the signatures of different egos. These results were consistent across different filters for the constructions of ego networks. Taken together, they suggest that in practice individuals allocate time and resources systematically when communicating. Regarding the Granovetter effect, we find that stronger ties are associated with overlapping circles of friends, a feature that persists even when considering different proxies for the strength of ties. While these results have been previously observed in contemporary datasets [26,33,29,9], historical datasets bring an added value on two fronts: (i) They provide evidence for human communication patterns in a distinct context—contemporary examples are usually the result of auto-recorded digital logs, and are thus representative of modern practices. (ii) They provide a timeframe that is unachievable in contemporary datasets, where samples of ego networks are examined across different decades, and where aggregate network evolution spans centuries.

5 Discussion

This paper presented new data resources and data services for historical epistolary linked open data. The presented CKCC and corresposearch datasets are, to the best of our knowledge, first LOD-based epistolary datasets available on the Semantic Web. Examples of analyzing and visualizing the data were presented and discussed using SPARQL querying and Python scripting as a proof-of-concept of the usability of the data resources. The aggregated data is now openly available for the research community for related analyses. We also demonstrated the idea of developing applications, i.e., semantic portals, of top the data service that require no programming skills from the end user. More details about the demonstrator can be found in [15] and there is an online video[35] demonstrating how the LetterSampo portal is used in practise on the Web, too.

This paper focused on presenting, discussing, and illustrating design principles for publishing and using epistolary data as linked data, not on presenting actual analysis results of particular datasets. This remains a topic of further research, but the first experiments presented show in our mind that the framework and the published resources, the linked open data and data service at LDF.fi, and the LetterSampo portals are promising in filtering our patterns of possibly interesting phenomena in Big Data using distant reading [25]. However, traditional close reading by a human is needed as before in interpreting the results.

A major challenge in creating data analyses like the ones shown in this paper is related to the quality of the data produced. Historical (meta)data is typically incomplete and our knowledge about it is uncertain. Also using more or less automatic means for transforming and linking the data leads to problems of incomplete, skewed, and erroneous data [23]. In historical epistolary data in particular, the data is seldom complete as only part of the letters have survived or are included in the data available. This as well as conceptual difficulties in modeling complex real world ontologies, such as historical geogazetteers, become sometimes embarrassingly visible when using and exposing the knowledge structures to end-users. In traditional systems the same problems are there, but are hidden in the non-structured presentations of the data. In general, more data literacy [22] is usually needed from the end-user when using data analytic tools.

The methods of network analysis can be very sensitive to even small errors in the data or biases in the sampling schemes. For example, betweenness centrality values can be dramatically changed by removal of even a single link, or long silences in communication in historical data can be explained by missing data from some historical period rather than inherently bursty communication tendencies. While computing various measures based on network data can be relatively simple with tools that are introduced here, the challenge that remains is correctly interpreting the results. This requires expert knowledge both in the domain to know how the data is biased and the methods to know how this affects the various measures. In the future, sampling schemes and missing data could be encoded in the data framework and the measures could be adopted to handle these situations. However, this work would first needed to be done within the domains (e.g., encoding sampling details of historical correspondence) and network method development (e.g., measures that consider missing data [21]).

Both of the two source datasets, CKCC and correspSearch, contained existing linkage to external LOD cloud databases which facilitated enriching the data by extracting, e.g., information about actor lifespans or geological metadata of places. Communication networks are easily huge, consisting of millions of links, which causes performance issues when, e.g., querying the database or rendering a large network on the web portal.

In spite of the challenges inherent in historical epistolary data, application of network analysis to the data can be useful for the researchers in finding out potentially interesting patterns of knowledge for closer study in datasets that are too big or complex for traditional manual means only. The new LOD resources and applications presented in this paper can now be used for this purpose. The framework is also being used in the

[35] `https://vimeo.com/461293952`

CoCo project[36] on Finnish correspondences in the Grand Duchy of Finland in the 19th century.

Acknowledgements We thank Arno Bosse, Howard Hotson, and Miranda Lewis for earlier collaborations during the *Cultures of Knowledge* project at the University of Oxford, funded by the Mellon Foundation, Charles van den Heuvel and Dirk van Miert for discussions related to CKCC, as well as colleagues in the EU COST Action project *Reassembling the Republic of Letters*[37]. Stefan Dumont provided the correspSearch data for our use. The work was also part of the *Open Science and Research Programme*[38], funded by the Ministry of Education and Culture of Finland, and the EU project InTaVia: In/Tangible European Heritage[39], and is related to the EU COST action Nexus Linguarum[40] on linguistic data science.

References

1. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the enron email corpus “it’s always about the people. enron is no different”. Computational & Mathematical Organization Theory **11**(3), 201–228 (2005)
2. Dumont, S.: correspSearch -– connecting scholarly editions of letters. Journal of the Text Encoding Initiative (10) (2016). https://doi.org/10.4000/jtei.1742
3. Eckmann, J.P., Moses, E., Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic. Proceedings of the National Academy of Sciences **101**(40), 14333–14337 (2004)
4. Goh, K.I., Barabási, A.L.: Burstiness and memory in complex systems. EPL (Europhysics Letters) **81**(4), 48002 (Jan 2008). https://doi.org/10.1209/0295-5075/81/48002, https://doi.org/10.1209/0295-5075/81/48002
5. Granovetter, M.S.: The strength of weak ties. American Journal of Sociology **78**(6), 1360–1380 (1973). https://doi.org/10.1086/225469, https://doi.org/10.1086/225469
6. Groth, P., Gil, Y.: Linked data for network science. In: Proceedings of the First International Conference on Linked Science - Volume 783. pp. 1–12. LISC’11, CEUR-WS.org, Aachen, DEU (2011)
7. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
8. van den Heuvel, C.: Mapping knowledge exchange in Early Modern Europe: Intellectual and technological geographies and network representations. International Journal of Humanities and Arts Computing **9**(1), 95–114 (3 2015). https://doi.org/10.3366/ijhac.2015.0140
9. Heydari, S., Roberts, S.G., Dunbar, R.I.M., Saramäki, J.: 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, https://doi.org/10.1007/s41109-018-0065-4
10. Hitzler, P.: A review of the semantic web field. Commun. ACM **64**(2), 76–83 (Jan 2021). https://doi.org/10.1145/3397512
11. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer–Verlag (2010)

[36] https://seco.cs.aalto.fi/projects/coco/

[37] http://www.republicofletters.net

[38] http://openscience.fi

[39] https://intavia.eu/

[40] https://nexuslinguarum.eu/the-action

12. Holme, P., Saramäki, J. (eds.): Temporal Network Theory. Springer–Verlag (2019)
13. Hotson, H., Wallnig, T. (eds.): Reassembling the Republic of Letters in the Digital Age. Göttingen University Press (2019), https://doi.org/10.17875/gup2019-1146
14. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012), https://doi.org/10.2200/S00452ED1V01Y201210WBE003
15. 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 (2021), https://seco.cs.aalto.fi/publications/2020/hyvonen-et-al-lettersampo-2020.pdf, submitted for peer review
16. Hyvönen, E.: ”Sampo” model and semantic portals for digital humanities on the semantic web. In: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 373–378. CEUR Workshop Proceedings, vol. 2612 (October 2020), http://ceur-ws.org/Vol-2612/poster1.pdf
17. Hyvönen, E.: Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web **11**(1), 187–193 (2020)
18. 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: Proceedings of the ESWC 2014 Demo and Poster Papers. Springer–Verlag (2014), https://doi.org/10.1007/978-3-319-11955-7_24
19. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web – Interoperability, Usability, Applicability (2021), accepted
20. Karsai, M., Kivelä, M., Pan, R.K., Kaski, K., Kertész, J., Barabási, A.L., Saramäki, J.: Small but slow world: How network topology and burstiness slow down spreading. Physical Review E **83**(2) (Feb 2011). https://doi.org/10.1103/physreve.83.025102, https://doi.org/10.1103/physreve.83.025102
21. Kivelä, M., Porter, M.A.: Estimating interevent time distributions from finite observation periods in communication networks. Physical Review E **92**(5), 052813 (2015)
22. Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science **49**(1), 3–14 (2015). https://doi.org/10.1177/0961000615616450
23. Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96. CEUR Workshop Proceedings (2020), http://ceur-ws.org/Vol-2612/paper6.pdf
24. van Miert, D.: What was the Republic of Letters? A brief introduction to a long history (1417–2008). Groniek **204/205**, 269–287 (2016)
25. Moretti, F.: Distant Reading. Verso Books (2013)
26. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences **104**(18), 7332–7336 (Apr 2007). https://doi.org/10.1073/pnas.0610245104, https://doi.org/10.1073/pnas.0610245104
27. Raji, P.S., Surendran, S.: Rdf approach on social network analysis. In: 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS). pp. 1–4 (2016). https://doi.org/10.1109/RAINS.2016.7764416
28. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web **8**(3), 373–383 (2017)

29. Saramaki, J., Leicht, E.A., Lopez, E., Roberts, S.G.B., Reed-Tsochas, F., Dunbar, R.I.M.: Persistence of social signatures in human communication. Proceedings of the National Academy of Sciences **111**(3), 942–947 (Jan 2014). https://doi.org/10.1073/pnas.1308540110, https://doi.org/10.1073/pnas.1308540110
30. Saramäki, J., Moro, E.: From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. The European Physical Journal B **88**(6) (2015). https://doi.org/10.1140/epjb/e2015-60106-6
31. Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., Hotson, H.: Reassembling the Republic of Letters - a linked data approach. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 76–88. CEUR Workshop Proceedings, vol. 2084 (March 2018), http://www.ceur-ws.org/Vol-2084/paper6.pdf
32. Ureña-Carrion, J., Leskinen, P., Tuominen, J., Hyvönen, E., Kivelä, M.: Communication now and then: Analyzing the Republic of Letters as a communication network (2021), https://arxiv.org/pdf/2112.04336.pdf, Submitted for publication.
33. Ureña-Carrion, J., Saramäki, J., Kivelä, M.: Estimating tie strength in social networks using temporal communication data. EPJ Data Science **9**(1) (Dec 2020). https://doi.org/10.1140/epjds/s13688-020-00256-5, https://doi.org/10.1140/epjds/s13688-020-00256-5
34. 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), https://arxiv.org/abs/2112.04336v1, in press
35. Vespignani, A.: Twenty years of network science. Nature **558**(7711), 528–529 (Jun 2018). https://doi.org/10.1038/d41586-018-05444-y, https://doi.org/10.1038/d41586-018-05444-y
36. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (Jun 1998). https://doi.org/10.1038/30918, https://doi.org/10.1038/30918
37. Wu, Y., Zhou, C., Xiao, J., Kurths, J., Schellnhuber, H.J.: Evidence for a bimodal distribution in human communication. Proceedings of the national academy of sciences **107**(44), 18803–18808 (2010)