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

Petri Leskinen [a,b,*], Javier Ureña-Carrion [c], Jouni Tuominen [a,b,d], Mikko Kivelä [c] and Eero Hyvönen [a,b]
[a] *Semantic Computing Research Group (SeCo), Aalto University, Finland*
E-mails: petri.leskinen@aalto.fi, jouni.tuominen@aalto.fi, eero.hyvonen@aalto.fi
[b] *HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland*
[c] *Complex Systems Group, Aalto University, Finland*
E-mails: javier.urenacarrion@aalto.fi, mikko.kivela@aalto.fi
[d] *HSSH – Helsinki Institute for Social Sciences and Humanities, University of Helsinki, 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. The key result of this paper is to present a framework, tools and systems, for creating, publishing, and analyzing historical LD from a network science perspective. The focus is on network analysis of epistolary network data (metadata about letters), based on recent advances in analysis of temporal communication networks and the behavioral patterns commonly found in them. To test, evaluate, and demonstrate the usability of the framework, it has been applied to (1) the Dutch CKCC corpus (of ca. 20 000 letters), (2) the pan-European correspSearch corpus (of ca. 135 000 letters), (3) to the Early Modern Letters Online data (of ca. 160 000 letters), and to (4) the aggregated Finnish CoCo collection of more than 300 000 letters from 1809–1917.

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

1. Introduction

Since the revolution in network science around 20 years ago [34, 40, 41], 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 allow for different levels of granularity ranging from 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.

[*] Corresponding author. E-mail: petri.leskinen@aalto.fi.

This paper builds on the idea that Semantic Web technologies[1] [12] and Linked Data [9, 16] 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 [9] can be used for making the data available for researchers in humanities with some skills on using SPARQL[2] 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 [19]. 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 created LD 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 [13] with practical LD publishing principles including SPARQL endpoints.

Table 1
Datasets analyzed and discussed in this paper

Dataset	Content
1. CKCC	Epistolary data of the CKCC corpus of the Huygens Institute in the Netherlands, an aggregated collection of ca 20 000 Dutch correspondences [10, 27] related to the Republic of Letters [15, 27]
2. correspSearch	Epistolary data 1510–1991 of 135 000 letters aggregated by the correspSearch project at the Berlin Brandenburg Academy of Sciences and Humanities [4]

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 [14, 34]. The datasets of historical epistolary data listed in Table 1 are used for 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 it has been found that the individuals are in contact in a bursty manner [6, 22] and they distribute their communication efforts via patterns, known as social signatures, that are specific to each individual [11, 33]. 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.

The paper extends our earlier papers related to publishing and analyzing historical epistolary data and LetterSampo [17, 37] by the network science perspective outlined above, and by presenting tools and systems for network analysis. The linked open data resources regarding datasets 1 and 2 of Table 1 are available online both as data dumps in Zenodo.org and in a SPARQL endpoint, as described in more detail in [17]. Several domain specific examples of using a demonstrator for epistolary research are presented in Section 4. and some more can be found in [17] and in an online video[3]. Using the data service for comparing epistolary network with modern communication neworks is discussed in [37]. 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

[1] https://www.w3.org/standards/semanticweb/
[2] https://www.w3.org/TR/sparql11-query/
[3] https://vimeo.com/461293952

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

2.1. 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 stage of Early Modern learned correspondences are proliferating rapidly, including, e.g., Europeana[4] [3], Kalliope Catalogue[5], The Catalogus Epistularum Neerlandicarum[6], Electronic Enlightenment[7], ePistolarium[8] [31], SKILLNET[9], correspSearch[10], the Mapping the Republic of Letters project[11], and Early Modern Letters Online (EMLO)[12] [10, 15, 27]. Visualizing the correspondences has been studied in the Mapping the Republic of Letters project[13] and in Tudor Networks of Power[14]. Bruneau et al. discuss applying Semantic Web Technologies to modelling the correspondences of French scientist Henri Poincaré and publishing on an online portal[15] [1].

The idea of representing epistolary data as a LD service was introduced in [36] and its application to DH research is discussed in [17] 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 in the four datasets listed in Table 1. 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 [19] and Sampo-UI framework [21], the data service can be integrated seamlessly with tooling for DH research making network analyses possible for researchers who often lack programming experience. (3) Thirdly, it is shown how the LD data service resource can be used for solving DH problems in network science with little programming experience using online programming services, such as Google CoLab[16] and Jupyter[17].

2.2. Temporal Network Analysis

In the past few decades, communication data has become a relevant resource to understand the underlying social networks [29, 34]. 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 epis-

[4] http://www.europeana.eu
[5] http://kalliope.staatsbibliothek-berlin.de
[6] http://picarta.pica.nl/DB=3.23/
[7] http://www.e-enlightenment.com
[8] http://ckcc.huygens.knaw.nl/epistolarium/
[9] https://skillnet.nl
[10] https://correspsearch.net
[11] http://republicofletters.stanford.edu
[12] http://emlo.bodleian.ox.ac.uk
[13] http://republicofletters.stanford.edu/
[14] http://tudornetworks.net/
[15] http://henripoincare.fr/s/correspondance/page/accueil
[16] https://colab.research.google.com
[17] https://jupyter.org

tolary 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 [37]. We identify two main approaches to analyzing such communication datasets according to the handling of temporality of the data [34, 37]. In a static approach a link is established between two people if there have been epistolary contacts between them, and in a temporal approach, the focus is on the distribution of dyadic interactions and behavioural features that characterize the way that people communicate. However, while most modern datasets attempt capture all auto-recorded communication within a communication channel (e.g., all emails or other communications within an organization [2, 5, 42]), this may not be true for historical data, since its collection is not automated, but implies broad manual compilation efforts by researchers.

For the static approach, a network is aggregated from dyadic interactions within a certain period or region. A link is created between two people if there has been some contact, and a proxy may be assigned for the strength of a tie based on, e.g., the total number of contacts [29]. From such static perspective it is possible to analyze 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 capture the relative importance of nodes within the network), or measures of the existence of communities or other types of structures.

For the temporal approach, a myriad of models have been proposed to analyzing network evolution [34]; we focus on the distribution of time-sequences of dyadic interactions, along with behavioural characteristics of how individual people communicate with their neighbours. From a sociological standpoint, the *Granovetter Effect* relates the notion of tie strength to network topology, noting that *strong ties* tend to be buried in overlapping circles of friends, akin to small communities where *weak ties* serve more as bridges between such communities [7, 29]. Since it is not possible to directly observe the strength of a tie, it is possible to use different temporal features as proxies [38]. Regarding the relationship of particular nodes to their neighbors, previous research [11, 33] has shown that individuals divide their contacting behaviour across their different neighbors in a persistent manner, known as a node's *social signature*, which is more stable in time than the neighbors themselves.

2.3. *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 [8] linked data is transformed for network analysis for the LinkedDataLens system. In [30] RDF data is used for Social Network Analysis. Data from different sources can be aggregated into larger networks and enriched by each other and by inferring 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 [9], 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; 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.

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

3.1. 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 the dataset-specific (*ckcc-schema*), *rdfs* refers to the RDF Schema[20], *crm* to the CIDOC CRM Schema[21], *geo* to WGS84 Geo Positioning vocabulary[22], *skos* to SKOS Simple Knowledge Organization System namespace[23], and *xsd* to the XML Schema of W3C[24].

The design choices are based on the principles developed in the EMLO project [36]. In the epistolary dataset, instances of the class *Actor* can be either people or groups. Each actor is connected to the sent letters using the property *:created* in a triple where the actor is the subject and the letter is the object. 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 these 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 and harmonized from different source formats. CKCC is an extract from an existing RDF dataset [36], while the correspSearch data was converted from a source published in the CMI format [4]. In these datasets both the actor and place resources had linkage to external LOD cloud databases, e.g., Wikidata, VIAF, Early Modern Letters Online project (EMLO), or database of Deutsche Nationalbibliothek[25] (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 were used to enrich our data, e.g., with images of actors and coordinates of the places. In our work, the "FAIR[26] guiding principles for scientific data management and stewardship" of publishing data are used.

3.2. 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* [17]. The SPARQL endpoint can be used directly in DH research using, e.g., YASGUI[27] [32] and Python scripting in Google Colab or 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[28] based on the Sampo model [19] for sharing collaboratively enriched linked open data using a shared ontology infrastructure. The portals host ready-to-

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

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

[22] http://www.w3.org/2003/01/geo/wgs84_pos#>

[23] http://www.w3.org/2004/02/skos/core#

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

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

[26] FAIR: Findable, Accessible, Interoperable, and Re-usable: https://www.go-fair.org/fair-principles/

[27] https://yasgui.triply.cc

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

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

use data-analytic tools for DH research, as suggested in [18]. The Sampo-UI framework [21] is used as the interface model and as the full stack JavaScript tool. Sampo portals are based—from a data perspective—on querying the SPARQL endpoint from the client side using JavaScript. The portals in the Sampo series 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.

3.3. *Querying and rendering networks in a web portal*

The networks in the portal pages are constructed using a customizable back-end service Sparql2GraphServer [25]. It was developed to meet the requirements for querying and constructing a network from any SPARQL endpoint. It builds a Sampo-UI compatible network based on SPARQL queries. It is a Python application built on Flask[29] framework using modules SPARQLWrapper[30] and NetworkX. The visual appearance of the network on a portal page is configured in the front-end Sampo-UI settings. The back-end service is used in other portals in the Sampo series like AcademySampo and ParliamentSampo. Figure 1 depicts a network extracted from Wikidata, it illustrates the teacher–student relationships starting from German polymath Gottfried Wilhelm Leibniz.

[29] https://flask.palletsprojects.com/en/2.2.x/

[30] https://sparqlwrapper.readthedocs.io

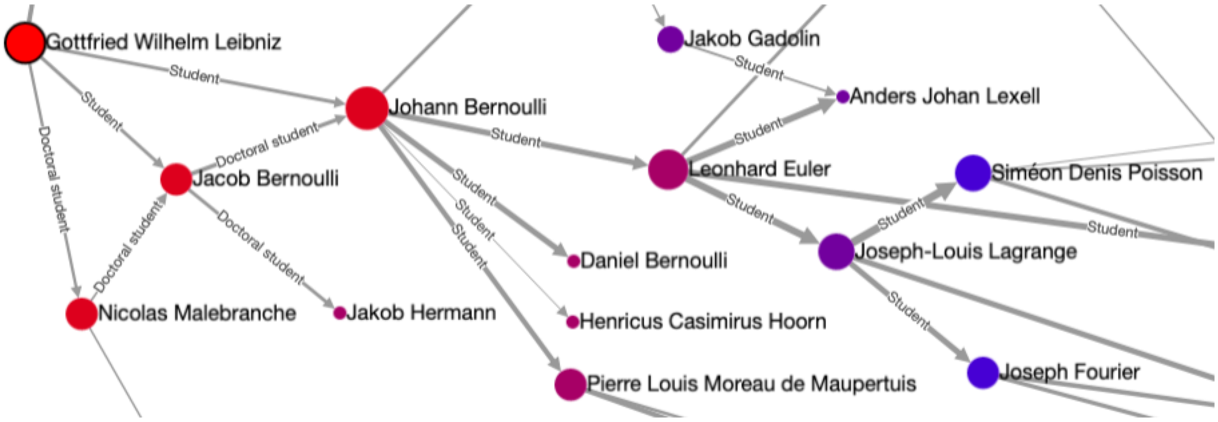


Figure 1. Social network of polymath Gottfried Wilhelm Leibniz in Wikidata

3.4. *New Resources on the Web*

The CKCC knowledge graph[31] as well as the correspSearch knowledge graph[32] have been published on the Linked Data Finland platform LDF.fi [20]. Both dataset are also available at Zenodo[33]. LDF.fi uses the *7-star* model for LD deployment [20] that extends the *5-star* model[34] 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 *seventh star* if validation results of the data using the schema are provided. LDF.fi is powered by the Fuseki SPARQL server[35] and Varnish Cache web application accelerator[36] for routing URIs, content negotiation, and caching. The portal user interface was implemented by the Sampo-UI framework [21]. The system uses Docker microservice architecture containers[37]. 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. The framework will be used in the *Constellations of Correspondence (CoCo)* project[38] on correspondences in the Grand Duchy of Finland in the 19th century [35].

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.

4.1. *Exporting Data for Data Analyses*

A 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 endpoint can be used. A benefit of using the

[31] The data, schema, and service are openly available (CC BY-NC 4.0) at the homepage https://www.ldf.fi/dataset/ckcc.
[32] The data, schema, and service available (CC BY-NC 4.0) at https://www.ldf.fi/dataset/corresp.
[33] CKCC: https://zenodo.org/record/6631385, correspSearch: https://zenodo.org/record/5972316
[34] https://5stardata.info/en/
[35] https://jena.apache.org/documentation/fuseki2/
[36] https://varnish-cache.org
[37] https://www.docker.com
[38] https://seco.cs.aalto.fi/projects/coco/

endpoint is that the data can be filtered and even transformed during the download to fit better for the aimed purpose. An example of using the data resource in external network analytic tools is presented in [39]. In this case study, the Linked Data of CKCC and correspSearch were analyzed 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 resemble each other strikingly even if the media were quite different.

4.2. General Analyses on Epistolary Networks

The knowledge graph also includes 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[39]. These measures are based on a network containing both the CKCC and the correspSearch datasets.

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 French philosopher and scientist René 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 socio-centric network visualization where the actors can be filtered, e.g., by their gender, years of living, or data sources.

4.3. 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 2. 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.

4.4. Using the LetterSampo Portal

Also a portal demonstrator[40] based on the aggregated CKCC and correspSearch LOD was published on the Web for public use [17]. The portal provides components for visualizing the epistolary data using line charts, maps, and networks. Figure 3 depicts an egocentric network around Descartes. In this visualization the widths of the edges are

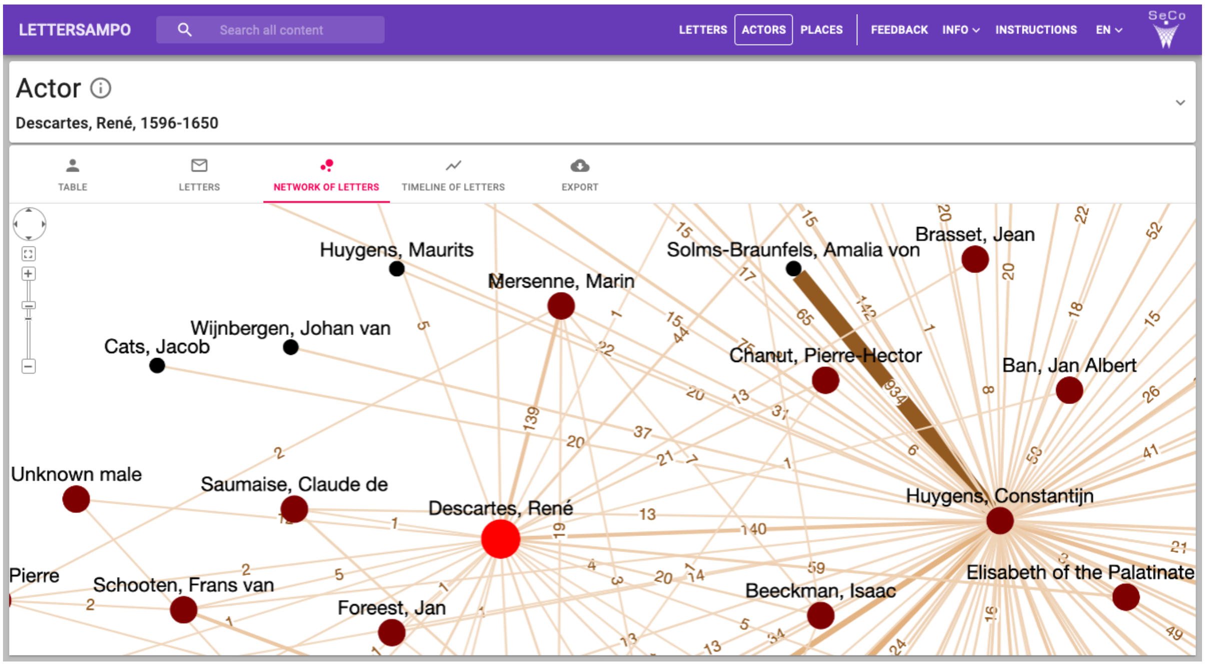
[39] https://networkx.org

[40] https://lettersampo.demo.seco.cs.aalto.fi/en

```
PREFIX lssc: <http://ldf.fi/schema/lssc/>
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 lssc:created ?letter ;
          skos:prefLabel ?source_label .

  ?letter a lssc:Letter ;
          lssc:was_addressed_to ?target ;
          skos:prefLabel ?letter_label ;
          lssc:has_time/skos:prefLabel ?date .

  ?target skos:prefLabel ?target_label .

} ORDER BY ?date
```

Figure 2. SPARQL example for querying the letters by René Descartes

proportional to the number of letters between the two actors while the sizes of the nodes are based on the length of the shortest path between the nodes 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, 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

Figure 3. Network of correspondences around René Descartes

Figure 4 depicts a visualization of the *social signatures* [11, 33] 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), a social signature is obtained by (1) computing the fraction of outgoing

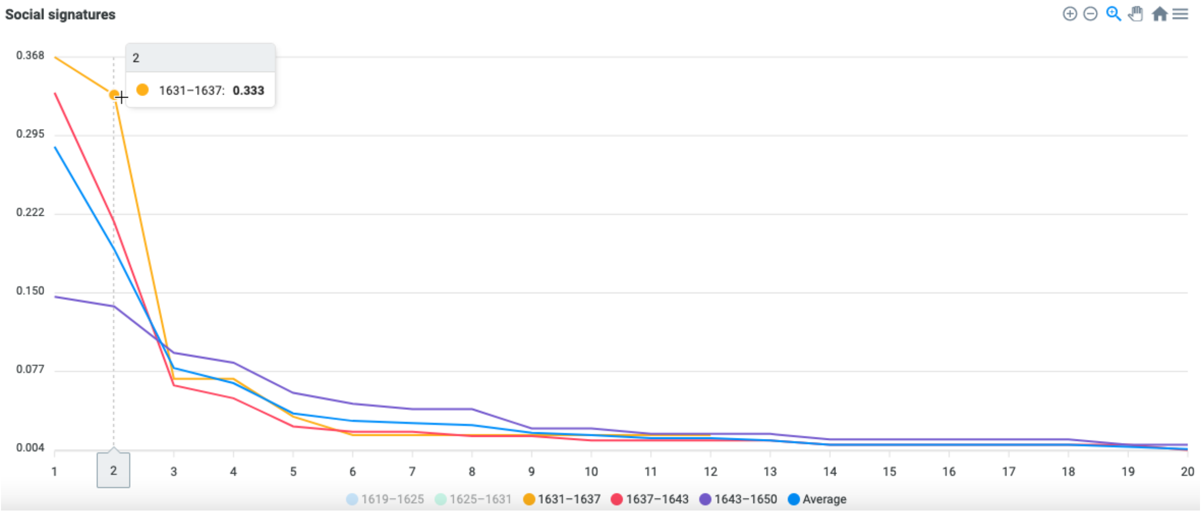


Figure 4. Chart depicting the social signatures of René Descartes

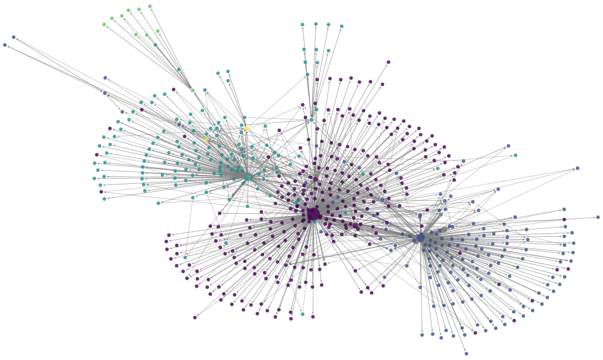


Figure 5. Network of CKCC data

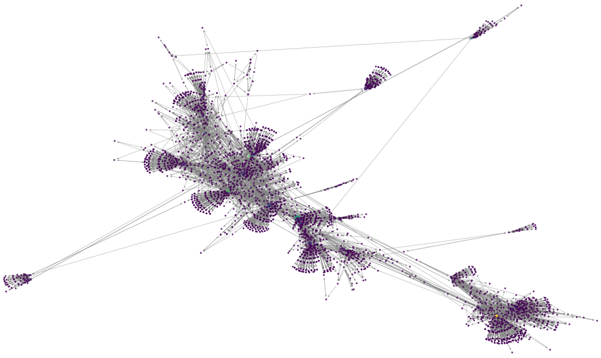


Figure 6. Network of correspSearch data

contacts per alter, and (2) ranking the alters. In the chart, like for instance the highest value of the yellow line is 0.368 indicating that Descartes wrote 36.8% of his correspondences to the top ranked alter, and likewise 33.3% to the second alter. This approach allows characterizing the relative importance of different alters in an ego network. When comparing different individuals, it is found that their social signatures tend to be stable [11, 33, 37].

4.5. Using the Endpoint by Programming

Due to the performance issues when attempting to render a larger network of more than, e.g., 1000 nodes on a browser page, data was further visualized in Google Colabs environment using Python. As an example, the largest connected component of the CKCC data is visualized in Figure 5. The network is built around three center actors: Dutch poet and composer Constantijn Huygens, philosopher Hugo de Groot, and mathematician and physicist 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 in Figure 6 has much more of these hubs.

Figure 7 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 French philosopher and mathematician Marin Mersenne, it can be observed that the source collections contain 134 letters from Descartes to Mersenne, but only five by Mersenne to Descartes.

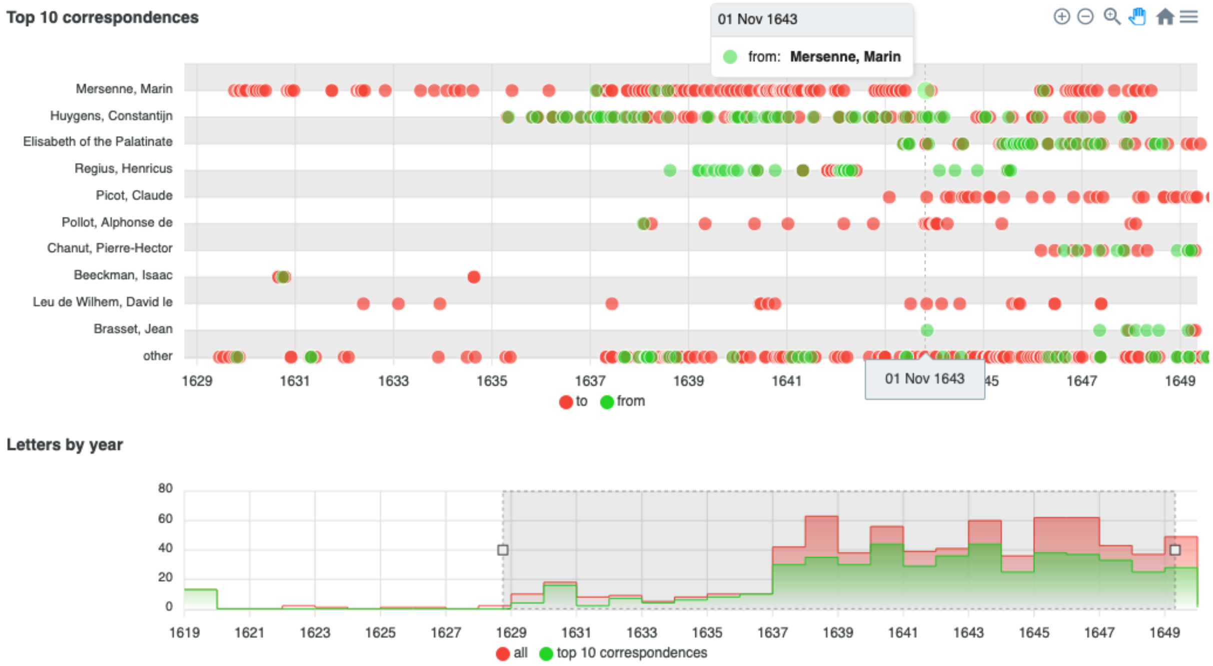


Figure 7. Timeline depicting top 10 letter correspondences of René Descartes

Figure 8 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 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. The code is available in GitHub[41] including a link to notebook in Google Colaboratory.

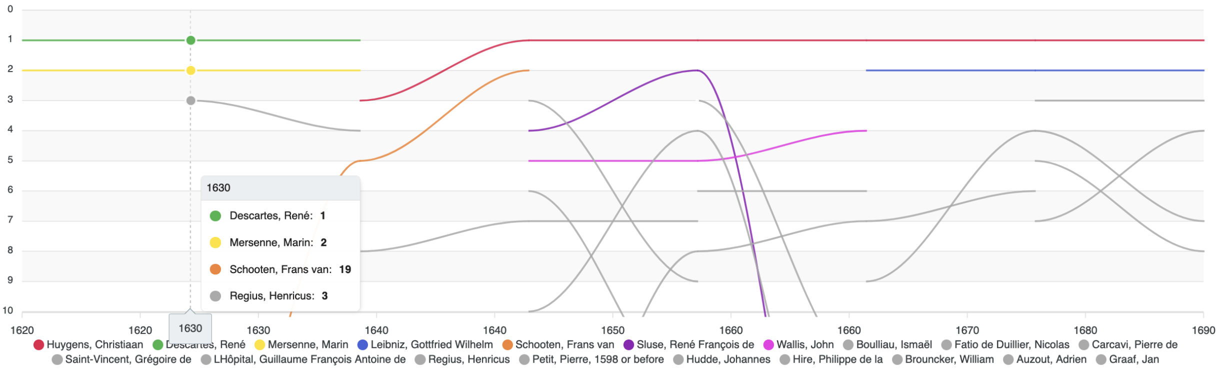


Figure 8. Top scientists in the CKCC data during 1620–1690

Figure 9 depicts the temporal evolution of the hubs in CKCC. In the figure each vertical rows on the x-axis corresponds to an actor and the years 1600–1700 are shown on the y-axis. The produce the image the correspondence

[41] https://github.com/SemanticComputing/LetterSampo-ckcc-charts

network of the 17th century was split into induced subgraphs each containing the correspondences during a time window of 2.5 years. From each subgraph twelve actors with highest total degrees are shown in the figure so that the one with the highest ranking has the brightest red color. In the figure the highest ranking actors Hugo de Groot, Constantijn Huygens, Christiaan Huygens, and Antoni van Leeuwenhoek stand out as the highest columns with red dots. One can notice that remain among the highest ranking ones during almost their entire time of floruit. On the other hand, 59.0% of people appear in the figure only once as a single blue dot. The code to produce this image is available in GitHub[42] including a link to notebook in Google Colaboratory.

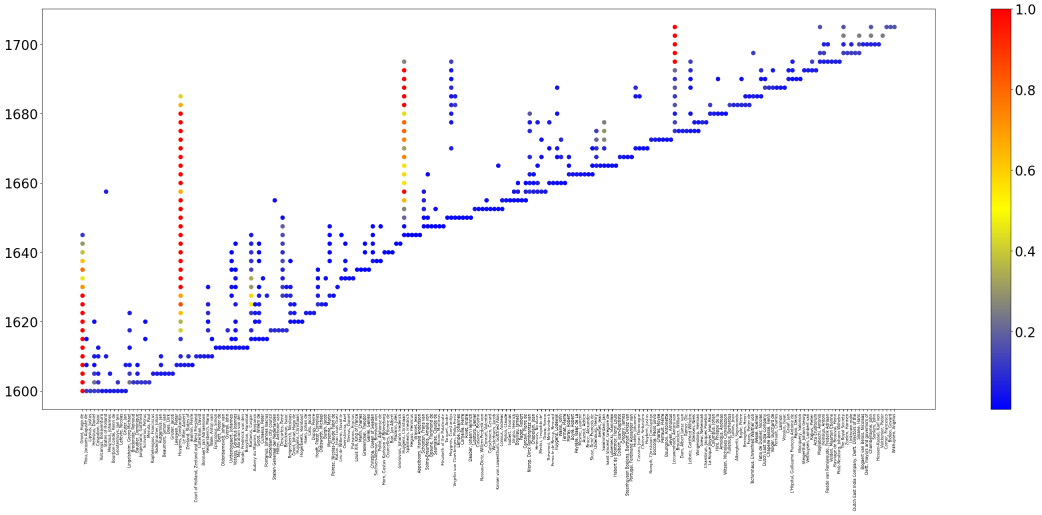


Figure 9. Evolution of ranking during the 17th Century in CKCC data

4.6. Comparing Contemporary and Historical Communication Networks

As a use-case scenario, a comparative analysis of historical and contemporary communication networks was performed, with results formally introduced in [37]. 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, the 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 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 and the data is biased.

Regarding social signatures, the 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, it is found 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 [11, 29, 33, 38], historical datasets

[42] https://github.com/SemanticComputing/LetterSampo-timeline

bring an added value on two fronts: (1) 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. (2) 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 tools and systems for analyzing networks of epistolary linked open data. Of the four datasets discussed, CKCC and correspSearch datasets are, to the best of our knowledge, first LOD-based epistolary datasets available on the Semantic Web. Examples of analyzing and visualizing the network 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 of these two datasets are openly available for the research community for related analyses. We also demonstrated the idea of developing applications, i.e., semantic portals, on top of the data service that require no programming skills from the end user.

This paper focused on presenting, discussing, and illustrating design principles for publishing and using epistolary network 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 opinion 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 [28]. 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 [26]. 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. The data is often also biased in different ways because historical data is often a result of a collection process performed by humans. For example, only letters of significant people have typically been collected in archives. It is therefore difficult to compare the underlying network with some modern networks, such as mobile phone networks, where the data has not been subject to human selection and is complete. This problem could be addressed by collecting data in unbiased ways or by trying analyze afterwards in what ways the data is biased.

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. The same problems exist in traditional systems but are hidden in the non-structured presentations of the data. In general, more data literacy [24] 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, the values of betweenness centrality can dramatically change 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 remaining challenge is to correctly interpret 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 need 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 [23]).

The datasets CKCC and correspSearch contained linkage to external LOD cloud databases which facilitated enriching the data by extracting, e.g., information about the lifespans of the actors 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.

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*[43]. Stefan Dumont provided the correspSearch data for our use. The work was also part of the *Open Science and Research Programme*[44], funded by the Ministry of Education and Culture of Finland, and the EU project InTaVia: In/Tangible European Heritage[45], and is related to the EU COST action Nexus Linguarum[46] on linguistic data science. CSC – IT Center for Science[47] provided computational resources.

References

[1] Bruneau, O., Lasolle, N., Lieber, J., Nauer, E., Pavlova, S., Rollet, L.: Applying and developing semantic web technologies for exploiting a corpus in history of science: The case study of the Henri Poincaré correspondence. Semantic Web **12**(2), 359–378 (2021)

[2] 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). , https://doi.org/10.1007/s10588-005-5377-0

[3] Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, p. 15. IFLA (2010)

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

[5] 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), https://doi.org/10.1073/pnas.0405728101

[6] 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

[7] Granovetter, M.S.: The Strength of Weak Ties. American Journal of Sociology **78**(6), 1360–1380 (1973). , htps://doi.org/10.1086/225469

[8] 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)

[9] 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/

[10] 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

[11] 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

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

[13] Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer–Verlag (2010), https://doi.org/10.1201/9781420090512-17

[14] Holme, P., Saramäki, J. (eds.): Temporal Network Theory. Springer–Verlag (2019), https://doi.org/10.1007/978-3-030-23495-9

[15] 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

[16] 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

[17] Hyvönen, E., Leskinen, P., Tuominen, J.: LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters. Journal on Computing and Cultural Heritage **16**(1) (2023), https://doi.org/10.1145/3569372

[18] Hyvönen, E.: 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), 187–193 (2020), https://doi.org/10.3233/SW-190386

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

[44] http://openscience.fi

[45] https://intavia.eu/

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

[47] https://www.csc.fi/en/home

[19] Hyvönen, E.: Digital humanities on the semantic web: Sampo model and portal series. Semantic Web – Interoperability, Usability, Applicability pp. 1–16 (2022), https://doi.org/10.3233/SW-223034

[20] 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

[21] 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 **13**(1), 69–84 (January 2022), https://doi.org/10.3233/SW-210428, online version published in 2021, print version in 2022

[22] 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

[23] Kivelä, M., Porter, M.A.: Estimating interevent time distributions from finite observation periods in communication networks. Physical Review E **92**(5), 052813 (2015), https://doi.org/10.1103/physreve.92.052813

[24] 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

[25] Leskinen, P., Hyvönen, E., Tuominen, J.: Sparql2GraphServer: a Server-side Tool for Extracting Networks from Linked Data for Data Analysis. In: ISWC-Posters-Demos-Industry 2021 International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. CEUR Workshop Proceedings (Oct 2021), http://ceur-ws.org/Vol-2980/paper343.pdf

[26] 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

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

[28] Moretti, F.: Distant Reading. Verso Books (2013), https://doi.org/10.1093/llc/fqu010

[29] 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

[30] 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

[31] Ravenek, W., van den Heuvel, C., Gerritsen, G.: The epistolarium: origins and techniques. CLARIN in the Low Countries pp. 317–323 (2017), https://doi.org/10.5334/bbi.26

[32] Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web **8**(3), 373–383 (2017), https://doi.org/10.3233/sw-150197

[33] 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

[34] 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

[35] Tuominen, J., Koho, M., Pikkanen, I., Drobac, S., Enqvist, J., Hyvönen, E., Mela, M.L., Leskinen, P., Paloposki, H.L., Rantala, H.: Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland. In: DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference. pp. 415–423. CEUR Workshop Proceedings, Vol. 3232 (March 2022), http://ceur-ws.org/Vol-3232/paper41.pdf

[36] 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

[37] Ureña-Carrion, J., Leskinen, P., Tuominen, J., Hyvönen, E., Kivelä, M.: Communication now and then: Analyzing the Republic of Letters as a communication network. Applied Network Science **7**(26) (2022), https://doi.org/10.1007/s41109-022-00463-1

[38] 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

[39] 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

[40] Vespignani, A.: Twenty years of network science. Nature **558**(7711), 528–529 (Jun 2018), https://doi.org/10.1038/d41586-018-05444-y

[41] 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

[42] 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), https://doi.org/10.1073/pnas.1013140107