Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland

Jouni Tuominen[1,2], Mikko Koho[2,1], Ilona Pikkanen[3], Senka Drobac[2], Johanna Enqvist[3,1], Eero Hyvönen[2,1], Matti La Mela[1,4], Petri Leskinen[1,2], Hanna-Leena Paloposki[3,5] and Heikki Rantala[2]

[1]*University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland*
[2]*Aalto University (Semantic Computing Research Group (SeCo)), Finland*
[3]*Finnish Literature Society, Finland*
[4]*Uppsala University, Sweden*
[5]*Finnish National Gallery, Finland*

Abstract

This paper presents the vision of aggregating, harmonizing, and publishing letter catalog metadata (information e.g. of senders, receivers and datings of letters) from cultural heritage (CH) institutions in Finland as a single reconciled Linked Open Data (LOD) service and a semantic portal providing data analytical tools for researchers. The research is conducted as part of the consortium research project *Constellations of Correspondence* (CoCo). The target of the project is to study – for the first time – scattered, heterogeneous epistolary metadata regarding the period of the Grand Duchy of Finland (1809–1917) as one, integrated dataset and make it interoperable and available. This will enable scholars to ask ambitious research questions in the field of computer science and to conduct empirical, bottom-up case studies e.g. on epistolary culture, communicative networks, and heritagization processes. This paper discusses one of the first datasets acquired by the project, the letter collection of the Board of the Finnish Art Society (1846–1901), provided by the Finnish National Gallery, which contains details of c. 1150 letters sent or received by c. 400 actors.

Keywords

Epistolary culture, letter metadata, Linked Open Data, semantic portal, data analysis, visualization

1. Introduction

Letters are essential sources for a wide variety of historical humanities. However, quantitative analyses are mostly absent even from the inquiries to epistolary cultures or letter-writing as a

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022

jouni.tuominen@helsinki.fi (J. Tuominen)
0000-0003-4789-5676 (J. Tuominen); 0000-0002-7373-9338 (M. Koho); 0000-0001-9435-7163 (I. Pikkanen); 0000-0002-7645-3079 (S. Drobac); 0000-0003-0901-7987 (J. Enqvist); 0000-0003-1695-5840 (E. Hyvönen); 0000-0003-0340-9269 (M. La Mela); 0000-0003-2327-6942 (P. Leskinen); 0000-0003-1412-8622 (H. Paloposki); 0000-0002-4716-6564 (H. Rantala)

social practice [1, 2]. The strong paradigm of close reading has rendered quantitative questions largely irrelevant. At the same time, as in the Finnish case, Cultural Heritage (CH) institutions (archives, museums, libraries) can rarely provide any reliable information as to the number of letters in their collections, which has also effectively prevented quantitative inquiries.

Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo) project's[1] main goal is to unite epistolary metadata of siloed collections and provide access to the harmonized, linked, and enriched dataset. A central output of the project will be an open data publication of aggregated letter catalog metadata that has been previously unavailable as a data resource. This will offer a new entry point to 19th-century epistolary culture, and it will enable both quantitative and qualitative analyses. Moreover, the project develops and utilizes social network analysis, data visualization, and knowledge discovery methods to respond to the empirical research questions. The analysis tools are packaged and provided for public and scholarly use as a LOD service, SPARQL endpoint, and semantic portal.

The project is currently surveying and collecting data from letter metadata collections scattered in different CH institutions in Finland. The collected datasets will be transformed into a harmonizing data model accompanied by automatic and manually curated disambiguation processes for the reconciliation of the identities of pivotal entities (people, places). The recognized people are linked to established LOD registers and enriched with their metadata, such as occupation and gender. An ontology of historical occupations [3] is used to link to information about social stratification provided via the HISCO standard [4]. For datasets with full-text contents of letters, we investigate options for enriching the metadata by utilizing natural language processing methods, such as named entity recognition. Possibilities of hand-written text recognition of digitized letters will be explored.

In the following sections, we present background on studying epistolary metadata collections, outline the currently available metadata collections we plan to use in the project, discuss initial ideas on the harmonized letter data model, data ingestion/aggregation workflow, and data publication, provide an initial demonstrator on a single letter metadata collection, and conclude with a general discussion.

2. What can epistolary metadata tell us?

The material aggregated during the project will enable us to ask ”big questions” regarding temporal patterns and trends of epistolary culture (How much correspondence as a whole and in chosen decades? Are there unexpected changes e.g. in the amount of letter exchange?). We can also utilize different visualization tools and quantitative measures e.g. from network science in order to find out who are the communicative hubs in different collections and in the 19th-century corpus as the whole. Are they persons we might expect in terms of corpus composition, or do the network metrics point beyond the ”usual suspects”, indicating need for further qualitative research and close-reading?

An equally important aspect is the study of the collections and the heritagization process related to their composition. The collections – and also their metadata – mirror what was

[1] https://coco.rahtiapp.fi

considered worth saving as heritage for the emerging nation and fit with the idea of the newborn Finnish national identity. As one result, we will acquire a more profound understanding regarding the ”data profiles” of different CH institutions (How do the explicit collection policies relate to the actual data they have accumulated?; Do the collections have specific temporal and gender profiles?). Such collection-related research is also a central part of the validation of the subsequent data models and visualizations and will form a backbone of source or ”data criticism” when the data is being used as a point of departure for humanistic inquiries.

The more specific research questions of this paper are connected to the letter collection of the Finnish Art Society (FAS) (in Finnish *Suomen Taideyhdistys*, STY), a predecessor of the present Finnish National Gallery and founded in 1846. Its role in the formation of the Finnish art world with all its institutions was crucial and it soon became the key actor in the 19th-century cultural field. The letters in this set were (and still are) attachments to the minutes of the Art Society, its board and other organs. In other words, they mainly contain issues that were discussed in the meetings. The set includes both letters sent by the Art Society and its representatives (copies) and letters sent to it. The minutes with the attachments and the letters are published online[2]. We will use the dataset of the FAS to briefly interrogate the usability of the visualizations provided by the semantic portal for art historical research. What kinds of trends and patterns can we observe in the data? Does it challenge or confirm the established image of the activities of the FAS?

3. Data

Datasets. In the first phase of metadata acquisition the project has collaborated with a selection of key CH institutions, including the Finnish National Gallery, Finnish Literature Society (SKS), Swedish Literature Society in Finland (SLS), National Archives, and National Library. Moreover, metadata about the correspondence of the ”national philosopher”, professor, and senator J. V. Snellman has been acquired from the Edita Publishing House. These institutions together hold central collections related to persons and institutions influential in the fields of art and literature, learning, science, and politics. The collections of SKS and SLS also include documents compiled by lower-class, uneducated people. In the course of 2022 and 2023, the project will expand its datasets to include other sectors of society (e.g. business and economics) and material from local museums and archives across Finland, which probably have much smaller but potentially interesting collections (Can we for example find 19th-century actors who wrote in minority languages?). Moreover, we are currently assessing the possibility of acquiring metadata from collections in private ownership. Such an inclusive data acquisition demands resources but is vital for the representativeness of the data.

The data comes in heterogeneous formats with different data granularities (from well-curated, individualized metadata to extensive correspondences between two persons merged into one metadata record). Most institutions have archival databases or systems, from where it is possible to acquire data in structured formats and as data dumps, but some only or partly maintain the epistolary metadata in MS Word format (see Table 1). In many cases, the project will be responsible for filtering out 19th-century metadata from a larger set of epistolary data.

[2] http://www.lahteilla.fi/styp/

Cultural heritage institutions and datasets			
Institution	Collection	Total collection size	Structured metadata
National Gallery	The Art Society	1147 letters	100% / csv
SLS	Albert Edelfelt	1310 letters	100% / csv
Edita/Seco	J.V. Snellman	1514 letters	100% / csv
SKS	Elias Lönnrot	6247 letters	100% / csv
National Gallery	Letter collections	c. 9900 letters	100% / csv
SLS	Letter collections	235 700 letters	100% / csv
SKS	Letter collections	c. 2000 archives	Database/MS Word
National Archives, Helsinki	Letter collections	420 archives / 200 000 files	Database/MS Word
National Library	c. 90 collections	?	MS Word

Table 1
Available source datasets as of February 2022.

Data model. As part of the project, a harmonizing data model for epistolary metadata collections will be developed. Compatibility with existing standards will be ensured, e.g., relevant CIDOC CRM classes and properties (e.g. *crm:E5_Event*, *crm:E21_Person*, *crm:E53_Place*, *CRM:E52_Time-Span*) will be used. Central classes of the data model will include Letter, Place (e.g. place of writing), Actor (e.g. sender, receiver), and Organization (e.g. host of collection), accompanied with classes representing information related to the archival processes of letters, e.g. Collection and Archival series. The core metadata of a letter includes sender, recipient, place of writing/sending, place of receiving, time of writing/sending, and data source.

Data ingestion/aggregation workflow. In the envisioned workflow for data processing, firstly, each source dataset is converted into a simple, uniform RDF format. In this format, all data fields in the source material are only converted into literal string-type values. The simple RDF format will be the same for all source data, independent of their original format, e.g., spreadsheet, Word document, or an extract from an existing database. Secondly, the produced simple RDF conversions are further harmonized into the applied CoCo data model. During this process the literal values are converted to resources according to the data model schema, e.g., actor names are split into family and given names, gender is statistically inferred from the given names, and timespan resources are generated based on the literal expressions of time. Next, actor and place resources are linked to external databases, e.g. Wikidata, GeoNames, BiographySampo [5], and AcademySampo [6]. Available information about an actor, such as given and family names, floruit, and possible geographical locations, is used for reconciliation. Actor resources are enriched with biographical information, such as the times and places of birth and death, name variations, vocations, known locations, inter-personal relations, and images. Place resources are enriched with label variations, position in place hierarchy, images, and geographical coordinates. Finally, the full CoCo dataset is assembled by reconciling all individual datasets constructed from different sources. Reconciling will be done both computationally as well as manually by domain experts. Furthermore, the data will be enriched by, e.g., calculating network statistics and constructing resources of correspondence between two actors.

Data publication and service. The letter metadata gathered in the project will be published in an open Linked Data service LDF.fi with a permissive license, according to the Linked Data

publishing principles and other best practices of W3C [7], including the provision of a public SPARQL endpoint[3], and FAIR principles[4]. The data service allows for building custom data analyses in Digital Humanities research using, e.g., YASGUI[5] [8] and Python scripting in Jupyter notebooks. The dataset will be accompanied with descriptive metadata enhancing its findability and possibilities for re-use. In addition, the data will be published as data dump, e.g. in Zenodo. By publishing the accumulated data as an open resource, new possibilities for distant and close reading of epistolary metadata of the Grand Duchy of Finland will be provided.

4. Semantic portal demonstrator

To perform preliminary studies on the data, one of the source datasets was converted into RDF using the LetterSampo framework [9]. The chosen dataset, *Finnish Art Society, Letters 1846–1901* data provided by the *Finnish National Gallery*, contains details of c. 1150 letters sent or received by c. 400 actors. The actor data consists of both cultural organizations and Finnish people with cultural significance, many of them having pages in, e.g., Finnish Wikipedia, AcademySampo, or BiographySampo. To enrich the chosen test data, person records were linked to Wikidata where c. 140 records were matched. The linkage was done by matching the person names and comparing the years or birth and death to the known years of activity in the correspondences. The data was enriched by importing details of e.g., birth, death, occupation, and person images.

The semantic portal demonstrator is based on the Sampo model [10] and is implemented using the Sampo-UI programming framework [11]. In the user interface, the faceted search for actor data allows the user to filter the results e.g. by a person's name, gender, or times of birth or death. As a result the user will acquire the possible image, full name, times of birth and death, and number of letters sent or received by that particular actor.

Two timelines of letters are depicted in Fig. 1. The upper one shows the 10 actors with the largest numbers or sent (green dot) and received (blue dot) letters. The actor with the largest number of letters both sent and received (Board of the FAS, *STY:n johtokunta*) appears at the top of the timeline. In comparison, the other actors have floruited for clearly limited times. The lower timeline in Fig. 1 illustrates the yearly amounts of letters in the test data set where it can be noticed that most of the letters are dated between 1875–1900. In Fig. 2, the places of sending the letters are visualized on a map.

5. Discussion

Domain knowledge representation, Semantic Web technologies, and shared ontologies, such as CIDOC CRM, provide a sustainable and collaborative cyberinfrastructure for pursuing humanist research [12]. The CoCo project builds upon the experiences accumulated during the Reassembling the Republic of Letters (RRL) [13], 1500–1800 (2014–2018) COST action and works in close collaboration with the LetterSampo initiative[6] [14, 9] that is building a

[3]https://www.w3.org/TR/sparql11-query/
[4]https://www.go-fair.org/fair-principles/
[5]https://yasgui.triply.cc
[6]https://seco.cs.aalto.fi/projects/rrl/

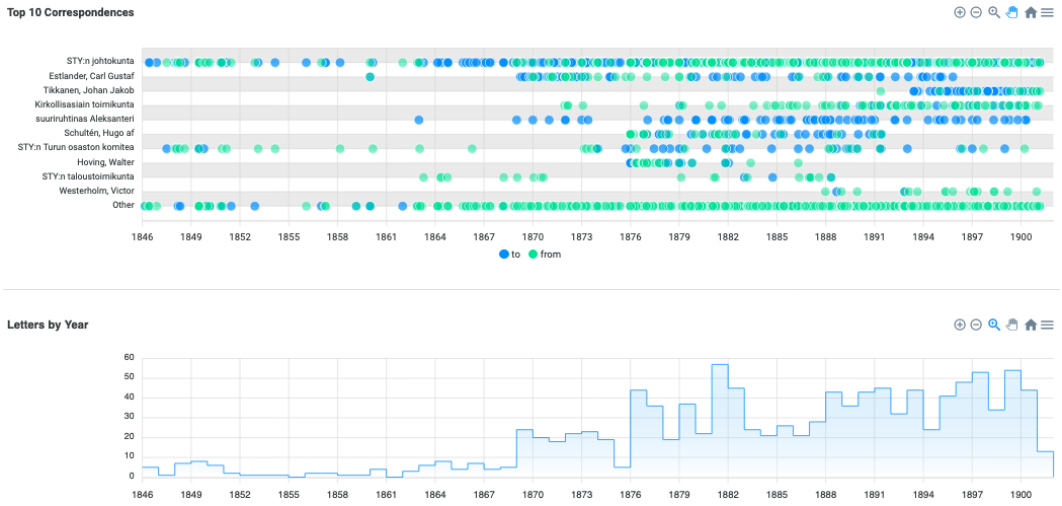

Figure 1: The upper chart depicts the most active actors and the lower chart the yearly amounts of letters.

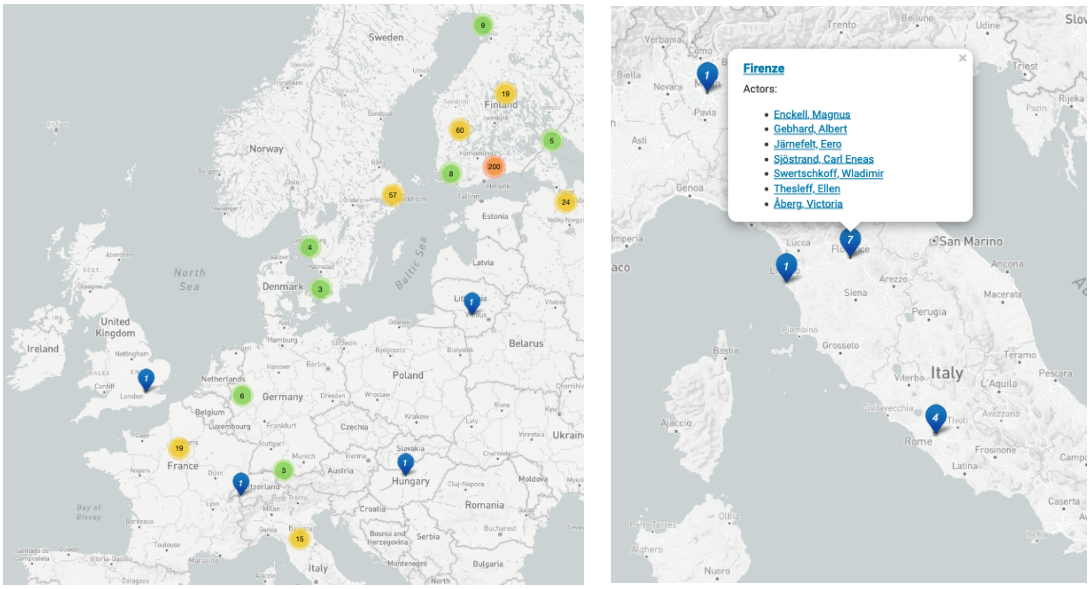

Figure 2: The places of sending the letters marked on the map of Europe (on the left) and Italy (on the right).

framework for representing, publishing, and using epistolary data as LOD on the Semantic Web for Digital Humanities research. The work utilizes our previous efforts on AMMO ontology of historical occupations, Bio CRM data model biographical information [15], Sampo-UI for building semantic portals, and Linked Data publications of various cultural heritage datasets (e.g. manuscripts, biographies). We collaborate with other correspondence metadata projects,

such as CKCC[7] [16, 17], correspSearch [18, 19], SKILLNET[8], the Early Modern Letters Online (EMLO)[9], and Norwegian Correspondences [20]. For modeling the textual contents of letters, there exists the work by the TEI Special Interest Group on Correspondence[10]. Although the CoCo project deals with Finnish data, the data model, reconciliation and enrichment workflow, and analysis tools developed will be applicable to other datasets as well, as one of the goals of the project is to produce a generic framework for producing LOD data publication based on heterogeneous epistolary metadata collections.

Compared to the many of the projects mentioned above, CoCo is focused on data that are both temporarily and geographically more restricted but, due to this, much more comprehensive and inclusive regarding 19th-century epistolary culture. A preliminary comparison between a larger dataset lately accumulated by CoCo (the metadata of approximately 300 000 letters; the more detailed discussion of this dataset is, however, beyond the scope of the present paper) and that of CKCC and correspSearch demonstrates, that there is a clear difference for example in gender balance (20 percentage of the authors are female in the Finnish case vs the mere 4 percentage in CKCC and correspSearch), indicating that female ”epistolary agency” will only become truly visible when the data are accumulated without a priori scholarly filters.

Regarding the Finnish Art Society's dataset, a researcher interested in the correspondence connected to the Society or its board can get a quick overview as to the main actors (senders) and the amounts of letters by simply studying the letter catalog in Excel format. However, even the quite limited and specialized data currently available can show the activities of the FAS from a fresh perspective and point to some of the fundamental source critical questions vis-à-vis the use of such data.

A large amount of prior knowledge is always at work when scholars interpret models. Also the list of 10 most active senders/receivers of letters mentioned above is not surprising. Persons such as Carl Gustaf Estlander or Johan Jacob Tikkanen are amongst the known leading figures of the FAS, and the Turku (Åbo) branch of the Society (STY:n Turun osaston komitea) a natural institutional correspondent. Perhaps the most interesting and important facet regarding the visualization of actors in this restricted dataset is that it brings the gendered nature of the Society's activities to the fore. Public organizations were run by men – and so was the Art Society, although for example its Drawing School was from the start open for both genders.

The epistolary metadata of the collection discussed here is very detailed, which means that altogether 823 entries contain information about the sending place. Put on a map, one of the places that stand out is Budapest, the location of the Hungarian Art Society. The societies e.g. discussed a possible regular exchange of Finnish and Hungarian lithographs. At the time, Finno-Hungarian connections were cherished particularly in Finland, since the Finns were set apart from Indo-European speakers as an ethnic group, and Finno-Ugrism became the favored ethno-linguistic alternative.

The number of the letters in the dataset increases throughout the century as the activities of the Society become more varied and frequent. The visualizations point to certain distinct

[7] CKCC is an acronym for “Circulation of Knowledge: A Web-based Humanities' Collaboratory on Correspondences and Learned Practices in the 17th century Dutch Republic ”.

[8] https://skillnet.nl

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

[10] https://tei-c.org/activities/sig/correspondence/

peaks in the correspondence dated to e.g. 1849, 1869 and 1876. Should we concentrate on these years, if we study the 19th-century activities of the FAS? On the other hand, is there a reason for a sharp reduction in the correspondence in 1875 or do we just observe casual fluctuation? Just to give one example: the peak in 1869 can be due to the revision of the statutes in 1868, which brought about an important change as the gathering of a permanent art collection was officially included amongst the responsibilities of the Society. It is a plausible hypothesis that such changes can cause accelerated epistolary activity.

At the same time, we should bear in mind the bounds of chance regarding the surviving of the data and, also, the archival practices of the Art Society and the later keeper of the collection, the National Gallery. It might well be that what we see in the data are the actions of a particular dutiful or perfunctory secretary. This may serve as a good reminder, that data-driven questions related to epistolary culture and inquiries as to the value-laden heritagization processes active in the formation and maintenance of the collections are more than two parallel research strands; they are deeply intertwined and should be taken into account on each level of the acquisition, processing and interpretation of epistolary metadata.

Acknowledgments. Our work was funded by the Academy of Finland as part of the project *Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo)* (decision numbers 339828, 340834, and 339918). CSC – IT Center for Science, Finland, provided computational resources for the work.

References

[1] D. Barton, N. Hall (Eds.), Letter Writing as a Social Practice, John Benjamins Publishing Company, 2000. doi:10.1075/swll.9.

[2] M. Leskelä-Kärki, A. Lahtinen, K. Vainio-Korhonen (Eds.), Kirjeet ja historiantutkimus, Suomalaisen Kirjallisuuden Seura, 2011.

[3] M. Koho, L. Gasbarra, J. Tuominen, H. Rantala, I. Jokipii, E. Hyvönen, AMMO Ontology of Finnish Historical Occupations, in: A. Poggi (Ed.), Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage, volume 2375 of *CEUR Workshop Proceedings*, 2019, pp. 91–96. URL: http://ceur-ws.org/Vol-2375/short2.pdf.

[4] M. H. van Leeuwen, Studying long-term changes in the economy and society using the HISCO family of occupational measures, in: Oxford Research Encyclopedia of Economics and Finance, Oxford University Press, 2020. doi:10.1093/acrefore/9780190625979.013.541.

[5] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research, in: The Semantic Web. ESWC 2019, Springer-Verlag, 2019, pp. 574–589. doi:10.1007/978-3-030-21348-0_37.

[6] P. Leskinen, E. Hyvönen, Linked open data service about historical Finnish academic people in 1640–1899, in: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, volume 2612 of *CEUR Workshop Proceedings*, 2020, pp. 284–292. URL: http://ceur-ws.org/Vol-2612/short14.pdf.

[7] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition), Morgan & Claypool, Palo Alto, California, 2011. URL: http://linkeddatabook.com.

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

[9] E. Hyvönen, P. Leskinen, J. Tuominen, LetterSampo – historical letters on the semantic web: A framework and its application to publishing and using epistolary data of the Republic of Letters, 2022. URL: https://seco.cs.aalto.fi/publications/2020/hyvonen-et-al-lettersampo-2020.pdf, submitted.

[10] E. Hyvönen, Digital Humanities on the Semantic Web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability (2022). URL: http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series-0, accepted.

[11] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, Semantic Web 13 (2022) 69–84. doi:10.3233/SW-210428.

[12] D. Oldman, M. Doeer, S. Gradmann, Zen and the art of Linked Data: new strategies for a Semantic Web of humanist knowledge, in: S. Schreibman, R. Siemens, J. Unsworth (Eds.), A New Companion to Digital Humanities, John Wiley and Sons, 2016, pp. 251–273. doi:10.1002/9781118680605.ch18.

[13] H. Hotson, T. Wallnig (Eds.), Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship, Göttingen University Press, 2019.

[14] J. Tuominen, E. Mäkelä, E. Hyvönen, A. Bosse, M. Lewis, H. Hotson, Reassembling the Republic of Letters – a linked data approach, in: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018), volume 2084 of *CEUR Workshop Proceedings*, 2018, pp. 76–88. URL: http://www.ceur-ws.org/Vol-2084/paper6.pdf.

[15] J. Tuominen, E. Hyvönen, P. Leskinen, Bio CRM: A data model for representing biographical data for prosopographical research, in: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017), volume 2119, CEUR Workshop Proceedings, 2018, pp. 59–66. URL: http://ceur-ws.org/Vol-2119/paper10.pdf.

[16] C. van den Heuvel, Mapping knowledge exchange in Early Modern Europe: Intellectual and technological geographies and network representations, International Journal of Humanities and Arts Computing 9 (2015) 95–114. doi:10.3366/ijhac.2015.0140.

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

[18] S. Dumont, S. Grabsch, J. Müller-Laackman, correspsearch – connect scholarly editions of correspondence (2.0.0) [web service], Berlin–Brandenburg Academy of Sciences and Humanities, 2021. URL: https://correspSearch.net.

[19] S. Dumont, correspSearch – connecting scholarly editions of letters, Journal of the Text Encoding Initiative (2016). doi:10.4000/jtei.1742.

[20] A. Rockenberger, E. N. Wiger, M. R. Witting, H. Bøe, E. I. Thor, O. J. Wolden, M. Paasche, O. Søndenå, P. Conzett, Norwegian correspondences and linked open data, in: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, volume 2364 of *CEUR Workshop Proceedings*, 2019, pp. 365–375. URL: http://ceur-ws.org/Vol-2364/33_paper.pdf.