Reassembling the Republic of Letters – A Linked Data Approach

Jouni Tuominen[1,2], Eetu Mäkelä[1,2], Eero Hyvönen[1,2], Arno Bosse[3], Miranda Lewis[3], and Howard Hotson[3]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
http://seco.cs.aalto.fi
firstname.lastname@aalto.fi
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://heldig.fi
[3] Faculty of History, University of Oxford, Oxford, UK
firstname.lastname@history.ox.ac.uk

Abstract. Between 1500 and 1800, a revolution in postal communication allowed ordinary men and women to scatter letters across and beyond Europe. This exchange helped knit together what contemporaries called the *respublica litteraria*, or Republic of Letters, a knowledge-based civil society, crucial to that era's intellectual breakthroughs, and formative of many modern European values and institutions. To enable effective Digital Humanities research on the epistolary data distributed in different countries and collections, metadata about the letters have been aggregated, harmonised, and provided for the research community through the Early Modern Letters Online (EMLO) catalogue. This paper discusses the idea and benefits of using Linked Data as the basis for a potential future framework for EMLO, and presents our experiences with a first demonstrator implementation of such a system.

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

1 Introduction

The revolution in postal communication in the early modern period allowed scholars and ordinary people to share their thoughts via letters in an efficient manner, in Europe and beyond. This development was a vital requirement for the *respublica litteraria*, or Republic of Letters, a knowledge-based civil society, crucial to that era's intellectual breakthroughs, and formative of many modern European values and institutions. However, for the modern scholars of the subject the scattered nature of the letter poses challenges, as the letter manuscripts are held in different libraries, archives, and private collections around the world.

Digital resources on early modern learned correspondence are proliferating rapidly but without a common framework for sharing data, tools, and systems development. Such resources include Europeana[4], Kalliope[5], The Catalogus Epistularum Neerlandi-

[4] http://www.europeana.eu
[5] http://kalliope.staatsbibliothek-berlin.de

carum[6], Electronic Enlightenment[7], ePistolarium[8], the Mapping the Republic of Letters project[9], and Early Modern Letters Online (EMLO)[10]. To reassemble the material and to facilitate its efficient study, coordinated discussions amongst librarians and archivists, scholars, IT and media experts are needed to collectively plan a shared digital infrastructure for publishing, reconciling, visualising, and analysing correspondence. Many of these conversations have been taking place over the last three years under the auspices of the EU COST Action IS1310 'Reassembling the Republic of Letters'[11].

This paper presents a Linked Data approach for such an infrastructure, using the Early Modern Letters Online (EMLO) collection as a pilot dataset. EMLO is a collaboratively populated union catalogue of sixteenth-, seventeenth-, and eighteenth-century letters, created by the Cultures of Knowledge project[12] at the University of Oxford. It brings manuscript, print, and electronic resources together in one space, increasing access to and awareness of them, and allows disparate and connected correspondences to be cross-searched, combined, analysed, and visualised.

The paper is organized as follows. First, the general vision and process description in our case study of creating, aggregating, and utilizing distributed epistolary data about letters is outlined, based on a Linked Data approach. After this, the underlying data models, data conversion, ontology services, tooling, and use of the data service in research are discussed.

2 A Distributed Publishing Model

Fig. 1 illustrates the overall process and setting considered in this paper. Epistolary data from different countries is being aggregated by the EMLO service (see the directed red arcs in the figure). The data can be accessed by the scholars using the portal. In our experiment, the legacy EMLO data was transformed into linked data, and published as a Linked Data Service in a SPARQL endpoint with additional services, such as content negotiation, linked data browsing etc. based on the W3C standards and best practices of Linked Data publishing [4]. In our experiment, the Linked Data Finland platform[13] [8], hosted by Aalto University was used[14] (see the blue arrow in the figure). Using Linked Data as a basis for aggregating and publishing the data has the following potential benefits for the overall process:

[6] http://picarta.pica.nl/DB=3.23/
[7] http://www.e-enlightenment.com
[8] http://ckcc.huygens.knaw.nl/epistolarium/
[9] http://republicofletters.stanford.edu
[10] http://emlo.bodleian.ox.ac.uk
[11] http://republicofletters.net
[12] http://www.culturesofknowledge.org
[13] http://ldf.fi
[14] Due to IP restrictions the data is currently not freely available, but access is being negotiated with the metadata owners.

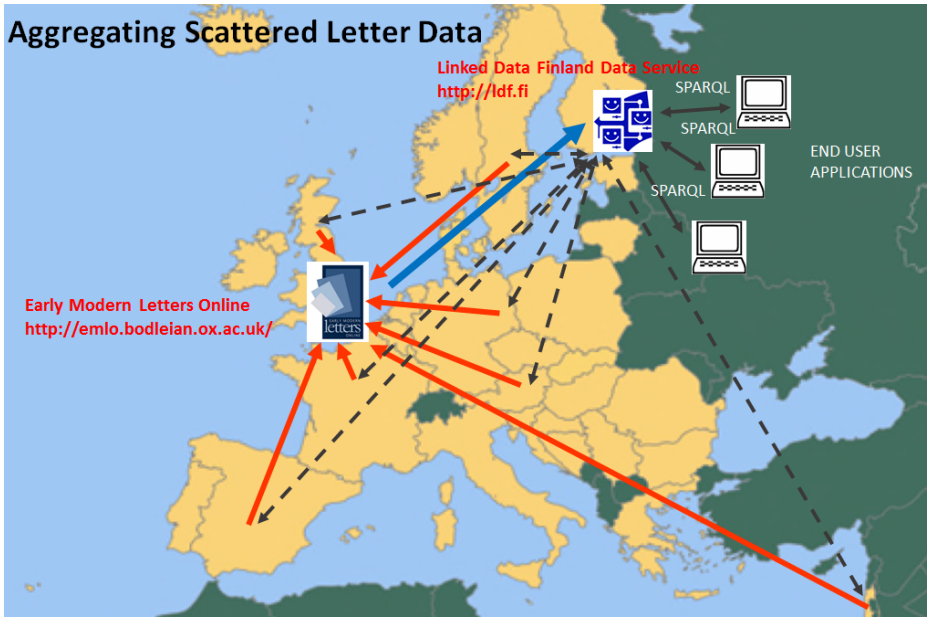

Fig. 1. Overview of the Linked Data approach in creating and aggregating distributed epistolary data.

1. *Data aggregation*. The RDF data model underlying the Semantic Web and Web of Data[15] is very flexible and simple for combining heterogeneous data from multiple data silos.
2. *Support for sharing ontologies*. Ontologies used in populating the metadata, such as historical people and places, can be shared within the community using ontology services [12].
3. *Crowdsourcing*. When cataloguing, new resources created in the distributed content creation network can be shared, as suggested in [7].
4. *Support tooling*. SPARQL endpoint provides a flexible standard API for creating tools for data cleaning, entity linking, ontology mapping, etc.
5. *Open application development*. In the same vein, the SPARQL API can be used in a standardized way for creating rich internet applications (RIA). No server side programming and data management is needed, if the API is available, which can simplify application development substantially and make it possible to virtually anyone.

The dashed arrows in Fig. 1 illustrate the fact, that the Linked Data service can be used not only in application development, but also during the data cataloguing process in the participating organizations. Using shared up-to-date ontology services, disambiguated identifiers for, e.g., persons and places can be assigned more easily and

[15] `http://www.w3.org/2013/data/`

duplication of work is avoided. Also tooling for, e.g., data cleaning, reconciliation, and duplicate checking can be shared in this way, saving human resources of the community as a whole and leading to more accurate and interoperable metadata from the outset.

3 Data Models and Linked Data Conversion

In order to allow scholars to efficiently study the vast amount of epistolary data from different data sources as a whole, the data has to be made semantically interoperable, either by mapping different data models (e.g., by using Dublin Core[16] and the Dumb-Down Principle[17]), or by providing a harmonised data model to transform the datasets into linked data [6]. We are suggesting the use of a shared data model for all the datasets. Unlike many other manuscript genres, letters share readily identifiable basic features (sender, recipient, date of sending and arrival, place of origin and destination) which facilitate the formation of a common data model.

In the context of EMLO, we have converted the original relational database via a straightforward conversion process using a script[18] into an RDF format. The conversion retains EMLO's internal data model, and thus follows a simple attribute-based modeling approach. A letter is represented as an instance of the class "Letter", and it has properties, such as "created" (inverse property), "was addressed to", "was sent from", "was sent to", "has time-span" (date), "original calendar", "language", "repository", "shelfmark", "printed edition details", and "source" (the catalogue the letter belongs to). The data model utilises CIDOC CRM[19] [2] (for time spans, people, and places), Dublin Core (for language, date, description, and subject), FOAF[20] (for person names and gender), and SKOS[21] (for labels) vocabularies.

In addition to purely epistolary data, EMLO contains prosopographical infomation related to the people in the database, modeled as events and social relationships. Events cover activities that the people have participated in during their lives, such as birth and death, ecclesiastic and educational activities, creations of works, travels and residences. The event metadata includes the event name, type, participants and their roles, time span, location, and source information. We converted the prosopographical data into RDF format using CIDOC CRM for the event-based modeling and W3C's PROV model [10] for representing the roles of participants in the events.

As a continuation of this work, we have also developed Bio CRM[22] [13], a semantic data model for harmonising and interlinking heterogeneous biographical information from different data sources. It is a domain specific extension of CIDOC CRM, effectively providing compatibility with other cultural heritage information as well. The data model includes structures for basic data of people, personal relations, professions,

[16] http://dublincore.org/documents/dcmi-terms/
[17] https://github.com/dcmi/repository/blob/master/mediawiki_wiki/Glossary/Dumb-Down_Principle.md
[18] http://github.com/jiemakel/anything2rdf
[19] http://cidoc-crm.org
[20] http://xmlns.com/foaf/spec/
[21] http://www.w3.org/TR/skos-reference/
[22] http://ldf.fi/schema/bioc/

and events with participants in different roles. One of the novelties of Bio CRM is the VIVO/BFO-inspired[23] [11], intuitive, and simple approach for the modeling of roles in different contexts – unitary roles, binary relationships, and events.

4 Ontologies and Ontology Services

For authority control, shared ontologies of people, places, and other relevant entity types, such as events, are needed. A natural starting point for creating such ontologies are the existing authority files, listings, and databases used in the data sources. In our use case, we converted the people and places used in EMLO into RDF format, using CIDOC CRM classes *E21_Person* and *E53_Place*. The idea is to store them in their own graphs in a public triple-store, where they can be queried and utilized by the community using SPARQL.

In cases where a data source uses a shared, established authority database, it can be used as such with a Linked Data approach. A number of authority sources such as VIAF[24], Getty ULAN[25], and CERL Thesaurus[26] already provide their data in RDF format, which further simplifies their utilisation.

For efficient use of the shared ontologies, we have developed the Federated SPARQL Search Widget[27], a user interface component that can be integrated into, e.g., letter cataloguing systems. Using such an approach, the different data providers already receive strong identifiers for the people and places as part of the data input process [1], with no need to reconcile the data later. Fig. 2 depicts an example of a SPARQL search widget for Finnish historical people, with contextual information supporting the selection of the correct person, including a person's photograph, short biographical description, and the places of activity visualised on a map.

5 Tooling for Reconciliation

When combining data from different sources, support tooling for reconciling the data into a harmonised format is needed. In the context of EMLO, there already exists a network of contributors – including scholars working on a specific collection or edition of correspondence, librarians, and publishers. These contributors provide metadata pertaining to the correspondence for ingestion into EMLO. The metadata can be input using a custom spreadsheet or via the EMLO-Collect online web form. Names of both authors and recipients (people), and origins and destinations (places) are included in the provided metadata. When inputting this data into EMLO, these people and places have to be matched to existing person and place records in the EMLO database or else assigned new person and place IDs. A semi-automatic tool, Recon[28], has been developed to assist with this matching process.

[23] http://vivoweb.org
[24] http://viaf.org/viaf/data/
[25] http://vocab.getty.edu
[26] http://www.cerl.org/resources/cerl_thesaurus/linkeddata
[27] http://github.com/SemanticComputing/federated-sparql-search-widget
[28] http://github.com/jiemakel/recon

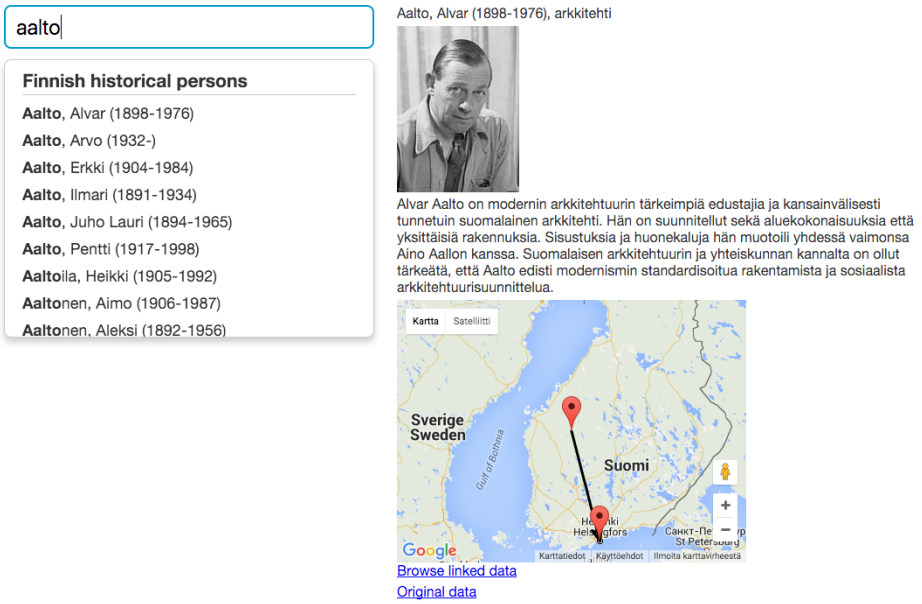

Fig. 2. A SPARQL autocompletion widget for Finnish historical people.

Recon is designed for digital humanities scenarios where trusted accuracy is of paramount importance. This means that: a) the matching cannot be done entirely automatically; b) the tool has to return as many potential matches as possible for the user to consult and consider a 'match'; and c) the user has to be supported in the manual verification process with the provision of contextual information concerning the match candidates. Compared to reconciliation tools such as Silk [15] and OpenRefine [14], Recon focuses on a manual review of potential match candidates, using a browser-based user interface to afford a simple, fast, and intuitive workflow.

The Recon user interface is depicted in Fig. 3. The tool reads a spreadsheet of names of people or places, possibly with contextual information, such as the years in which a person was active (*floriat*). Working through the data rows, Recon runs SPARQL queries to a triple-store containing people and places extracted from the current EMLO database. For each person or place in the spreadsheet, a list of potential candidate matches is offered to the user, based on the string similarity of the name, and potentially other criteria based on the SPARQL query used in the matching process. For example, the years of activity of a person can be used to rank candidates with suitable birth and death years higher than those similarly named people who have lived at some other time period. The user has the option to specify whether there is a match or not, or to leave a query open in case there is an uncertainty; this query might request further investigation be carried out. When the spreadsheet has been processed, Recon re-exports to the user the original data supplemented with the EMLO IDs of the matched people or places. Where no matches have been identified, new EMLO records are created and

their IDs inserted. Following this, the revised dataset can be ingested into EMLO using this complete list of people and place IDs.

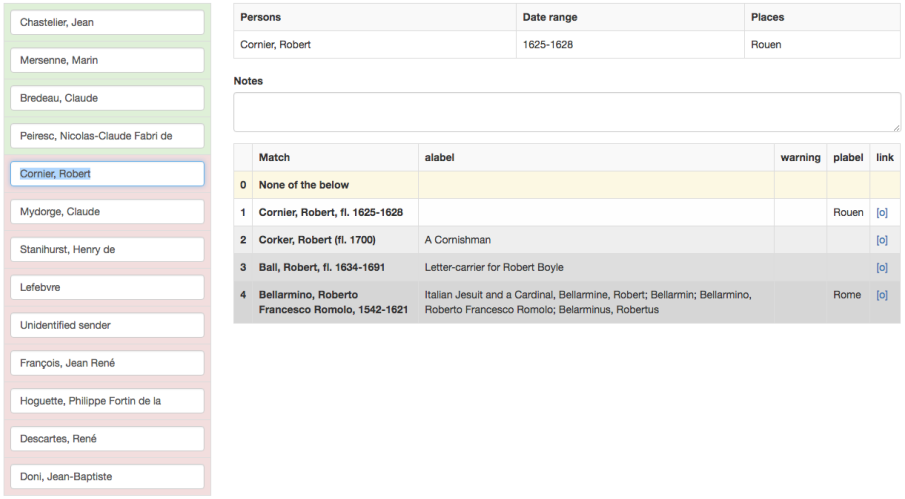

Fig. 3. The user interface of Recon.

For pre-processing tabular letter metadata into a more efficient format before Recon is used, a complementary tool called Mare[29] has been developed. Mare is a map/reduce user interface for tables. The tool is used in the EMLO spreadsheet workflow to collect all unique people and place names from a correspondence dataset with contextualizing information, such as the years of activity based on the dates of the letters that involve particular people or places. A sample output of Mare is depicted in Fig. 4.

In addition to using Recon for the semi-automated matching of newly contributed datasets, the tool has been piloted to enable the identification and linking of records for the same letters contained in separate catalogues within EMLO. To achieve this, Recon is configured to run SPARQL queries across the EMLO dataset to identify potential 'matching' letters, i.e., letters that have the same sender and recipient, and share similar or exact data in other metadata fields, in particular repository and shelfmark references, or printed edition details, dates, and places of origin and destination. The tool ranks the potential duplicate matches for a given letter by taking into account the proximity of the dates, string similarities of textual metadata fields, etc. The EMLO editors are then able to assess whether the entries provided by different contributors in different letter collections (whether they be listings of an early modern individual's correspondence or of a thematic collection) refer to the same letter; if the same letter has been entered by different contributors, a bridge link between the two 'interpretations' of the same letter can be inserted in EMLO.

[29] `http://github.com/jiemakel/mare`

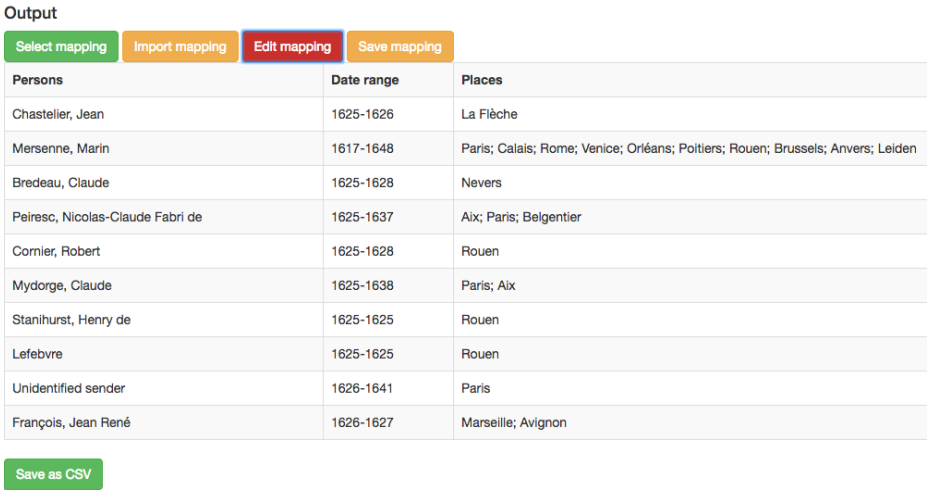

Fig. 4. A sample output of the Mare tool listing unique people, their activity years, and places involved in a letter catalogue.

Whilst working with Recon, EMLO's editors are able to call up records to identify matches allowing them to review people, place, and letter records in different combinations and to view the correspondence metadata 'from different angles'. In consequence, errors are spotted and corrected more easily, as well as partial matches, and can be cleaned and augmented in tandem, as appropriate.

6 Visualisation and Analysis Tools

The epistolary data published in a structured format can be conveniently visualised using general-purpose data visualisation and exploration tools, such as Palladio[30] [3], RAW[31], or SPARQL Faceter [9]. Palladio can not only ingest data from a spreadsheet, but the data can also be loaded directly from a SPARQL endpoint. This allows for the creation of live visualisations without the need to export data manually each time. Palladio can be used, e.g., for graph, timeline, or map-based visualisations. SPARQL Faceter allows a scholar to interactively examine a dataset by filtering it using different facets, such as sender, recipient, origin, destination, date, or catalogue.

Fig. 5 visualises the temporal distribution of the catalogues included in EMLO, using the RAW data visualization framework. One can see that EMLO contains different catalogues (colour-coded) of letters from the time period 1500–1800, with the highest peak representing correspondence activity in the 1640's. Fig. 6 visualises the social relationships of Samuel Hartlib based on the prosopographical data in EMLO

[30] http://hdlab.stanford.edu/palladio/
[31] http://app.rawgraphs.io

(connections of two steps from Hartlib), using Palladio. The map shows the connections Hartlib had to various locations around Europe (the size of a circle represents the amount of connections), and from the timeline one can see, e.g., that Hartlib was most active in the 1640's. Further visualizations of Hartlib's network using extended prosopographical data not yet integrated into EMLO may be viewed and queried via a pilot Shiny/R dashboard[32].

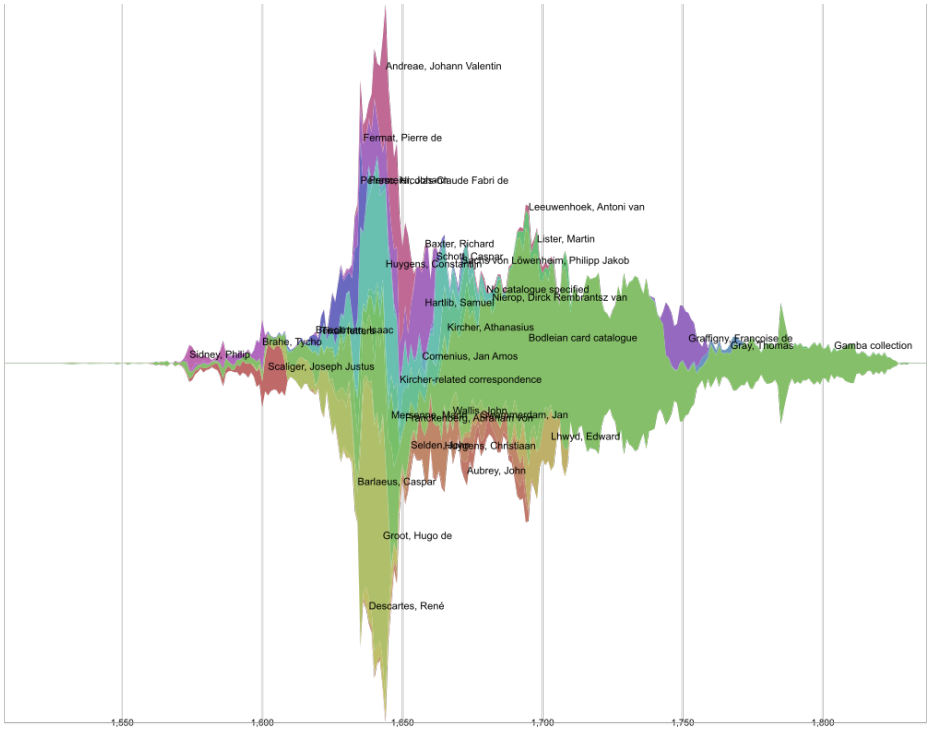

Fig. 5. The temporal distribution of the correspondence in the catalogues in EMLO.

7 Discussion

This paper presented the idea of using Linked Data as a basis for aggregating, harmonising, publishing, and using epistolary data in a distributed setting. To test and demonstrate the ideas, the existing EMLO service data was re-used, transformed into Linked Data, and published as a "5-star"[33] Linked Data service. On top of the SPARQL endpoint provided by the data service, further tools were created which could be utilised

[32] https://idn.web.ox.ac.uk/article/cultures-knowledge-case-study
[33] http://5stardata.info

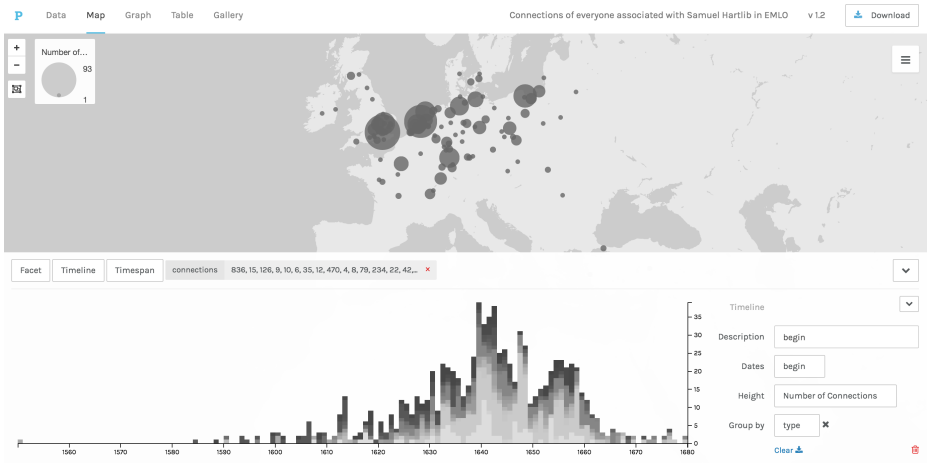

Fig. 6. Samuel Hartlib's social relationships visualised on a map and timeline.

by the scholarly community. The Mare and Recon tools are already in active use by EMLO's editors at the University of Oxford. We also demonstrated the potential of application development on top of the linked data service, by using Palladio and RAW for visualising the epistolary data from a digital humanities research perspective.

This paper focused on epistolary data only, but the Republic of Letters is of course not only about letters, but scholarly communications and the exchange of knowledge more broadly, including books, essays, artifacts, etc. A major benefit of the Linked Data approach in the future is that the model is flexible enough for representing different kind of forms of scholarly and cultural heritage content in an interoperable, machine "understandable" (semantic) way, including both tangible and intangible aspects of culture and history [6]. Based on semantic representations of knowledge, new kind of services based on, e.g., intelligent data analysis, Artificial Intelligence, and Knowledge Discovery can be conceived and created.

However, the envisioned potential and benefits also have a price tag. Legacy systems already in use do not yet support Linked Data, and the technology is new and not consistently established in IT departments. The most important challenge is, however, that using the new model requires greater collaboration and mutual agreements between the participating organizations, which complicates the process. One has to take into consideration the shared ontologies and vocabularies used by the community, not only one's own preferred standards and practices. However, since in this case the final goal of the community is to create a global view of the Republic of Letters, it is a better idea to avoid interoperability problems before they arise by a Linked Data infrastructure than to try to solve them afterwards when the damage is already done [5]. As Alfred Einstein put it: *Intellectuals solve problems, geniuses prevent them.*

Acknowledgements Our work is part of the EU COST Action project *Reassembling the Republic of Letters*[34] and the *Cultures of Knowledge* project, funded by The Andrew W. Mellon Foundation. The work is also part of the *Open Science and Research Programme*[35], funded by the Ministry of Education and Culture of Finland.

References

1. Andert, M., Berger, F., Molitor, P., Ritter, J.: An optimized platform for capturing metadata of historical correspondence. Digital Scholarship in the Humanities 30(4), 471–480 (2015), https://doi.org/10.1093/llc/fqu027
2. Doerr, M.: The CIDOC CRM—an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003), https://doi.org/10.1609/aimag.v24i3.1720
3. Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., Coleman, N.: Historical research in a digital age: Reflections from the mapping the republic of letters project. The American Historical Review 122(2), 400–424 (2017), https://doi.org/10.1093/ahr/122.2.400
4. 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/
5. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic Web Journal 1(1–2), 33–37 (December 2010), http://www.semantic-web-journal.net/content/preventing-interoperability-problems-instead-solving-them
6. 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
7. Hyvönen, E., Tuominen, J., Ikkala, E., Mäkelä, E.: Ontology services based on crowdsourcing: Case national gazetteer of historical places. In: Proceedings of the ISWC 2015 Posters & Demonstrations Track. CEUR-WS Proceedings (2015), http://www.ceur-ws.org/Vol-1486/paper_45.pdf, vol 1486
8. 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
9. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf, vol 1615
10. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology (2013), http://www.w3.org/TR/2013/REC-prov-o-20130430/, W3C Recommendation 30 April 2013
11. Smith, B., Almeida, M., Bona, J., Brochhausen, M., Ceusters, W., Courtot, M., Dipert, R., Goldfain, A., Grenon, P., Hastings, J., Hogan, W., Jacuzzo, L., Johansson, I., Mungall, C., Natale, D., Neuhaus, F., Overton, J., Petosa, A., Rovetto, R., Ruttenberg, A., Ressler, M., Rudniki, R., Seppälä, S., Schulz, S., Zheng, J.: Basic formal ontology 2.0 – specification and

[34] http://www.republicofletters.net
[35] http://openscience.fi

user's guide (2015), https://github.com/BFO-ontology/BFO/raw/master/docs/bfo2-reference/BFO2-Reference.pdf, June 26
12. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009). pp. 768–780. Springer–Verlag (2009), https://doi.org/10.1007/978-3-642-02121-3_56
13. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Biographical Data in a Digital World (BD2017) (2017), https://doi.org/10.5281/zenodo.1040712
14. Verborgh, R., De Wilde, M.: Using OpenRefine. Packt Publishing (2013)
15. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Proceedings of the 8th International Semantic Web Conference (ISWC 2009). pp. 650–665. Springer–Verlag (2009), https://doi.org/10.1007/978-3-642-04930-9_41