Analyzing Aggregated Knowledge Graphs on a Global Level for Better Data Literacy: Case LetterSampo Finland

Henna Poikkimäki[1 [0000−0003−3362−8438]],
Petri Leskinen[1 [0000−0003−2327−6942]], and
Eero Hyvönen[1,2 [0000−0003−1695−5840]]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland

Abstract. Epistolary letter collections are stored in distributed local archives as letters are sent from one place to another. To find and study letters of a particular person or group on a global level, data from different local sources can be aggregated and harmonized into a global knowledge graph (KG). This paper argues that it is important to understand possible quality issues of the global KG that may arise due to the heterogeneity of the local datasets, aggregation process, and mutual linkedness of the local data. For example: In what ways do the local collections enrich each other? How complete is the aggregated dataset? Are there duplicates or misaligned entities and concepts in the aggregate? This paper presents a set of data-analytic tools to address such issues in order to support data literacy in Digital Humanities (DH) research. As a case study, the LetterSampo Finland Linked Open Data (LOD) KG is considered and the results are reported. It aggregates data about almost 1.3 million historical letters sent in the Grand Duchy of Finland (1809–1917) and 118 000 related actors harvested from 18 different archival data sources and 1670 fonds, enriched by data from 12 external databases.

Keywords: knowledge graphs, linked data, digital humanities

1 Introduction

Correspondence through letters has been an important means of communicating knowledge, opinions, and personal affairs since the invention of postal services and the rise of the Republic of Letters 1500–1800 [13, 6]. Letters that have survived to this day are usually kept in collections of different memory organizations, i.e., archives, libraries, museums, and galleries. Archival fonds typically focus on the correspondences of one person, but may also include a wider range of people, such as families. Collections are often biased, focusing on people considered prominent at the time when letters were collected.

More and more often, letter collections are digitized to varying degrees to provide better access to humanities researchers to study. Sometimes only metadata related to letters and related people and organizations are digitized, and

sometimes data might be enriched using different methods. For example, if the contents of the letters are available, the people and places mentioned in the letters can be recognized manually or by using the named entity recognition (NER) and linking (NEL) to external data sources [12].

The division of letters into heterogeneous, geographically distributed local collections and the varying quality of the digitized data makes it difficult to study communications on a global level. For example, one can study the personal correspondence network of some individual based on a single collection [17], but the individual's position in the larger communication network might not be evident. In general, combining datasets, especially atypical ones, tends to increase the scientific impact of the resulting publication [22].

As a solution for combining letter collections, the use of LOD has been proposed and used [21, 7] for aggregation, harmonization, and publication of letters as KGs. This approach was used in the LetterSampo Finland portal[3] [8] whose KG is available as LOD in Zenodo[4] and as a SPARQL endpoint on the LDF.fi platform[5]. The availability of such aggregate collections also facilitates novel analyses at the collection level. For example, to what extent do the collections enrich each other and provide mutually conflicting information?

Combining multiple collections into one KG comes with some difficulties. In aggregated collections, there might be letter duplicates, and it might be difficult to link individuals to themselves in other collections due to different spellings of the name, unclear birth and death years and places, and so on. Differences in metadata quality and missing letters hinder data analyses. Understanding the characteristics and limitations of the data is a prerequisite for reliable data analysis [11]. In this paper, we focus on the following research questions relevant for DH research based on aggregated data from a data literacy [10] point of view.

1. What kind of data quality issues arise in aggregated KGs?
2. How to make data quality issues related to data aggregation transparent to the end user to support data literacy?
3. How to find out how aggregated local collections enrich each other on a global level, i.e., determine the added value for aggregating data?

We use ontological data models and Linked Data to address these questions with a set of tools based on a SPARQL endpoint and Jupyter notebook scripting. As a case study, the LetterSampo Finland KG LOD service is used [8].

2 Related Work

Various archives have stored letter collections for future generations to study. The problem with using letter collections in DH research is that the data are distributed in different cultural heritage organizations, and the data have to be aggregated and then harmonized. Harmonizing letter metadata is challenging from

[3]LetterSampo portal: `https://kirjesampo.fi`

[4]LetterSampo KG: `https://zenodo.org/records/15210590`

[5]SPARQL endpoint of the LetterSampo KG: `https://www.ldf.fi/dataset/coco`

a technical perspective, as letters in different collections may have been written in different languages and cataloged using different data models and vocabularies. After aggregation and harmonization, heterogeneous data have to be provided to the research community through databases or web services. Examples of such services include Europeana[6], Kalliope[7], The Catalogus Epistularum Neerlandicarum[8], Electronic Enlightenment[9], ePistolarium[10], the Mapping the Republic of Letters project[11], SKILLNET[12], correspSearch[13], and the Early Modern Letters Online (EMLO) catalogue[14].

After data harmonization, the varying accuracy and coverage of the metadata in different letter collections can cause problems. Data are considered high quality when they are fit for use by consumers [19]. In the case of LetterSampo Finland data, there are two main purposes: 1) to query the letters and 2) to do analysis based on the metadata. In both cases the availability and quality of the letter and actor metadata is essential. One of the characteristics of historical data is the fragmentation of the data and missing data; i.e. many letters have gone missing for various reasons, the letters have not been digitized or the digitization is lacking. For example, the effects of missing data on the network metrics for historical social networks have been studied in [3] and networks based on letter collections with different types of missing data in [18].

Linking the same person in different data sources, or even within one data source to the same entity, is a common problem in linked data. Interlinked cultural heritage linked open datasets are not as interconnected as one would expect and links between datasets are not often reciprocal, making it more difficult for the data users to travel through knowledge graphs (KGs) [20]. Overall, linked data or the data on the Web suffers from large variation in data quality between data sources [23].

The interlinking of the resource can be studied by calculating network metrics based on the local neighborhood of the resource, by checking if there are open chains formed by the *owl:sameAs* or similar predicates, and how much new information is added to the resource through these predicates [5]. The connectivity between data sets has been studied through common entities, triples, literals, and schema elements [14]. The linked data quality metrics are often classified into four dimensions: accessibility, intrinsic, contextual and representational [15], and they help to evaluate the data from different perspectives. Several metrics for evaluating semantic accuracy, completeness, conciseness, consistency and syntactic validity, as well as metrics to study other linked data dimensions, are presented in [1] and [23].

[6] `http://www.europeana.eu`
[7] `http://kalliope.staatsbibliothek-berlin.de`
[8] `http://picarta.pica.nl/DB=3.23/`
[9] `http://www.e-enlightenment.com`
[10] `http://ckcc.huygens.knaw.nl/epistolarium/`
[11] `http://republicofletters.stanford.edu`
[12] `https://skillnet.nl`
[13] `https://correspsearch.net`
[14] `http://emlo.bodleian.ox.ac.uk`

In our case, we focus on the aggregated LetterSampo Finland KG based on the data in 18 different data sources with varying data quality. We want to see the quality of the metadata in different letter collections and consider consequences for the data analyses. Most importantly, we want to study how combining letters from multiple sources affects the personal networks of actors, which is why we chose script-based approach instead of existing schema languages like Shapes Constraint Language (SHACL)[15] or Shape Expressions (ShEx)[16].

3 LetterSampo Finland KG and LOD Service

LetterSampo Ontology Design Fig. 1 shows the data model behind LetterSampo Finland with its shared ontology infrastructure, including e.g. the CIDOC Conceptual Reference Model (CRM), Simple Knowledge Organization System (SKOS) and Dublin Core. The most important classes of the model are *:Letter* for modeling letter data and *:ProvidedActor* for actor data. Depending on the letters themselves and the digitization process, each letter has a varying amount of metadata available. Common metadata fields are language, data source, and the place and date of sending. Each actor with different labels in each data source is presented as CIDOC CRM class *crm:E39_Actor*. In general, there are four actor types: person, family (*:Family*), organization (*crm:E74_Group*), and unknown (*:Unknown*) which are all subclasses of *crm:E39_Actor*.

There can be multiple instances of *crm:E39_Actor* for one actor across the data sources with different labels that are connected to one *:ProvidedActor* instance that combines all available information about the actor, such as birth and death years, type, gender, and occupation if the actor is a person. How well linking *crm:E39_Actor* instances to correct *:ProvidedActor* instance has succeeded is critical to data analyses. The *:ProvidedActor* instances also have possible links to external sources through the *owl:sameAs* predicate. The most important external sources include Wikidata, AcademySampo containing academic records and BiographySampo containing biographies of Finnish people.

Harvesting and aggregating the LOD Currently, the LetterSampo Finland LOD service contains metadata for almost 1.3 million letters and related actors from 18 archival data sources that host 1670 fonds, received through questionnaires to Finnish archives. A tedious cleaning process and pipelines, described in [2], were needed for LOD transformation during which several challenges arose. The data were in various heterogeneous forms that often needed human interpretation, and there were issues with data quality, errors, and incomplete data. Linking and aligning actors to correct unique *:ProvidedActor* class entities was a major challenge as person names change in time due to, e.g., marriages and deliberate name changes, or names can have different spellings. To address the problem actor metadata, e.g., the times and places for birth and death, the known name variations of individuals, family relations, and occupations assembled from external sources were utilized.

[15] `https://www.w3.org/TR/shacl/`

[16] `https://shex.io/`

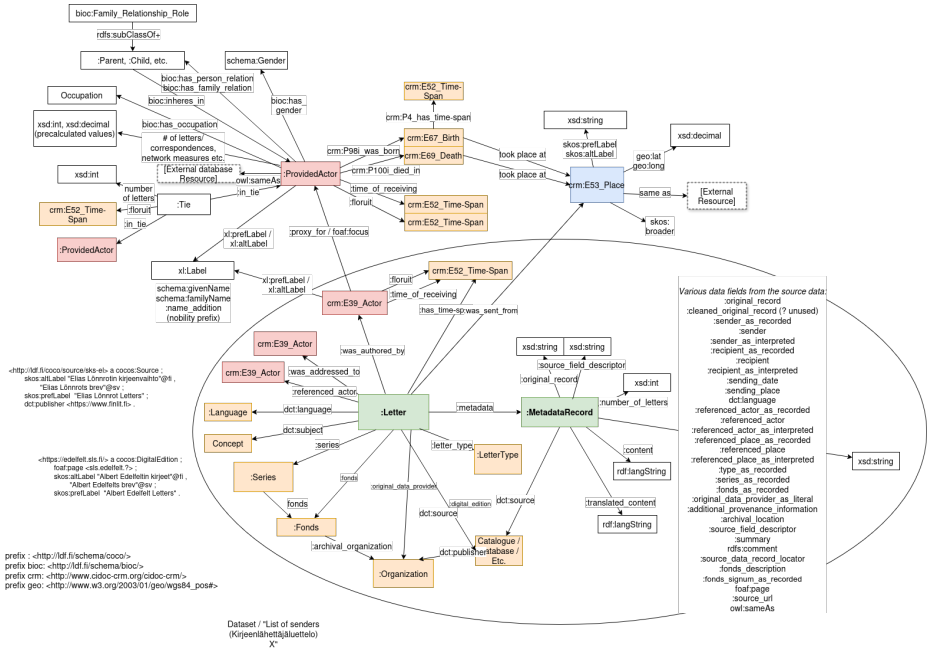

Fig. 1: The data model behind LetterSampo FinlandThere are two major classes: *:ProvidedActor* that handles the actor metadata and *:Letter* that handles the letter metadata. The default schema domain namespace is *(http:/ldf.fi/schema/coco/)*.

From a data perspective, a major challenge in the case study was that in many, if not most cases, letter-wise metadata were not available, but only metadata about archival units. For example, a particular unit in an archive may contain n letters that two families exchanged during a time period t, but it is not known who sent what letter to whom and what was the exact sending date or year. On the other hand, in some cases pertaining to people of national importance, very detailed metadata about individual letters, including content annotated with mark-up such as TEI[17] was available. Another challenge of the data is its massive size: the KG contains information about over 1.2 million letters and their related entities.

Based on the harmonized data, the LOD service and SPARQL endpoint using the Linked Data Finland platform LDF.fi[18] [9] was established as part of the national FIN-CLARIAH research infrastructure [19]. The LOD service SPARQL API can be used directly for DH research by, e.g., the Yasgui SPARQL query

[17] Text Encoding Initiative TEI: `https://www.tei-c.org/`

[18] Linked Data Finland platform: `https://ldf.fi`

[19] FIN-CLARIAH/DARIAH-FI: `https://seco.cs.aalto.fi/projects/fin-clariah/`

editor or Jupyter Notebooks. Some examples of using network analysis on epistolary data, using the whole data set and the egocentric network based on the correspondences of the polymath Elias Lönnrot are presented in [16].

4 Implementation

In the following sections, we present different issues that arise when aggregating local epistolary datasets and show how such issues, in a selection of cases, can be made transparent to the end-user for better data literacy. As a case study and example, the LetterSampo Finland KG is used, but most of the methods and tools implemented are generic and can also be adapted to datasets in other application domains.

The implemented tools start by querying the data from a SPARQL endpoint and saving the query results into tables. In our case study, we query actors and related metadata, and letters and related metadata. For analyzing the changes in networks, senders and receivers of the letters and related metadata, like the data source of the letters. The resulting dataframes from the queries are saved in parquet files[20] for further analysis.

Jupyter Notebooks and Python scripting are used to visualize and analyze potential problems in the data. In most cases, by giving a related dataframe and chosen column names for functions, one can get table-formatted results or visualize the data in helpful ways, e.g., the availability of the metadata for any dataframe. The Jupyter Notebooks and related scripts with more detailed documentation of the methods and tools are available on GitHub[21].

5 Enriching Data: Connections between Collections

We consider simple sender-receiver networks in order to see how combining multiple data sources affects actors' personal networks. In these networks, nodes are actors. When one actor has sent a letter to another, the directed edge goes from sender to receiver. Each edge has a weight that corresponds to the number of letters sent. The in-degree of actor is the sum of the weights of in-coming links, that is, the number of letters actor has received, and the out-degree is the number of letters actor has sent. The degree centrality of the actor is the sum of in-degree and out-degree, and an actor with a high degree can usually be considered to be a notable person within the network. The neighborhood of the actor are the other actors with whom the actor has been in direct contact. We consider the actor and their neighborhood as the personal network of the actor. Well-known actors are usually present in multiple epistolary data sources [18]. We expect combining data sources to broaden especially their personal networks.

[20] Apache Parquet format: `https://parquet.apache.org/`

[21] Github repository: `https://github.com/SemanticComputing/coco-about-data.git`

Fig. 2 shows how the actors are shared between the data sources. Edelfelt and Snellman Letters have a relatively high number of actors compared to the number of letters. This is because they also have letter contents available and contain actors mentioned in the letters which leads also to larger number of unique actors in the source as the mentioned people include older historical actors like Platon or Martin Luther and other people. The mentions to the other actors in the letters can be used to enrich actor's personal social network, and the contents of the letters can give further context for the links through close reading [4]. The five largest data sources in terms of the number of actors, including Åbo Akademi University Library, the National Archives of Finland, the National Library of Finland, the Society of Swedish Literature in Finland (SLS) and Finnish Literature Society (SKS), contain 98% of the letters.

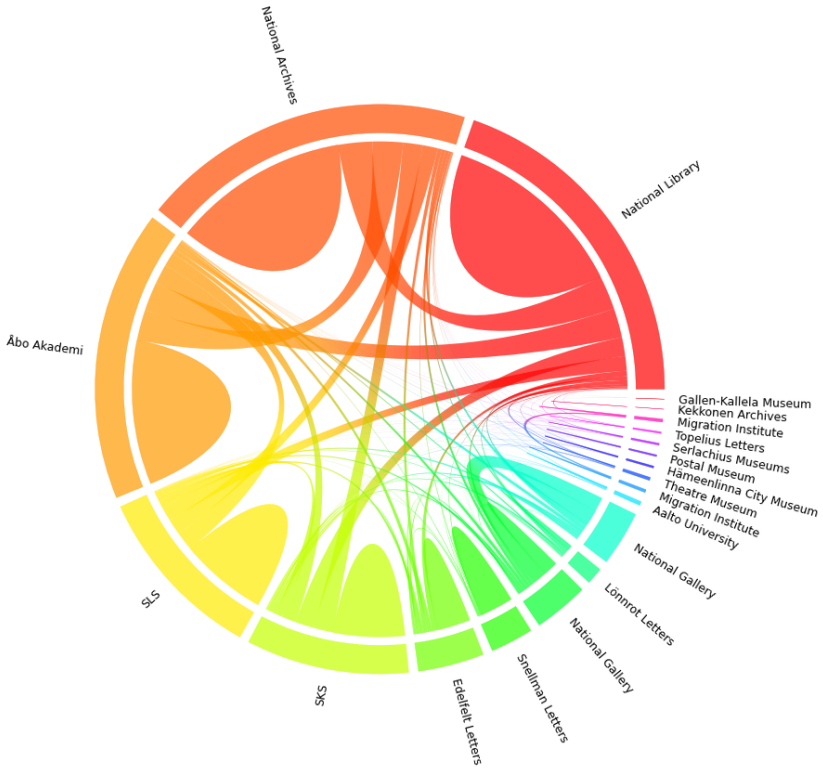

Fig. 2: Shared and unique actors per source. Snellman, Edelfelt, and National Gallery Letters contain also actors mentioned in the letters.

For five largest data sources the proportion of actors unique to the source is over 70%. Most of the actors belong only to one data source and also have a low degree on average, as seen in the Fig. 3. As the number of data sources an actor

belongs to grows, so does the average degree. When the mentioned actors are not taken into account, the actors, mainly migrants, in the Migration Institute of Finland data are the second most separated from the actors in other data sources as 91% of the actors are unique to the source and all 10 actors in the Archives of President Urho Kekkonen do not appear in any other data source.

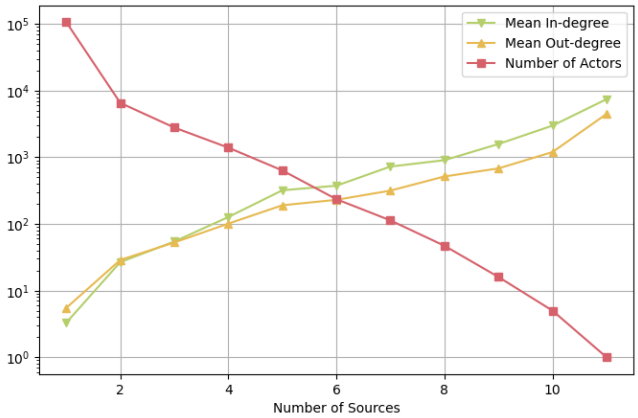

Fig. 3: Mean degree(s) of actors vs the number of data sources they are found in, and number of actors who are present in corresponding number of data sources.

Fig. 4 shows how much the neighborhood size grows compared to the "primary source" where actor has the largest number of neighbors, when all data sources actor appears in are combined. Letters sent and received are considered separately. Approximately one tenth of all actors belong to multiple sources, and 60% of them associated with at least 10 letter are considered here. In general, letters sent increase the size of the neighborhood more than letters received. Combining data sources helps to get better understanding especially about to whom actors has sent letters. Letters received by an actor are more likely to end up in one collection, whereas letters sent are stored in collections of multiple people. Out of all actors present in multiple data sources, approximately 60% have only sent letters but not received any. For some of the actors, the size of the neighborhood can grow manyfold when all data sources are included compared to the primary source.

For some actors, the most important correspondences based on the number of letters exchanged are family members or colleagues (e.g. Albert Edelfelt) and those letters are stored in one data source, whereas letters to other people are distributed to other sources. For those actors, including only the sources with the largest number of letters can result in a very diminished network. On the other hand, when a new actor is added to the neighborhood based on letter or two, we can say that there is definitely a connection between the two, but defining the importance of the connection requires close reading.

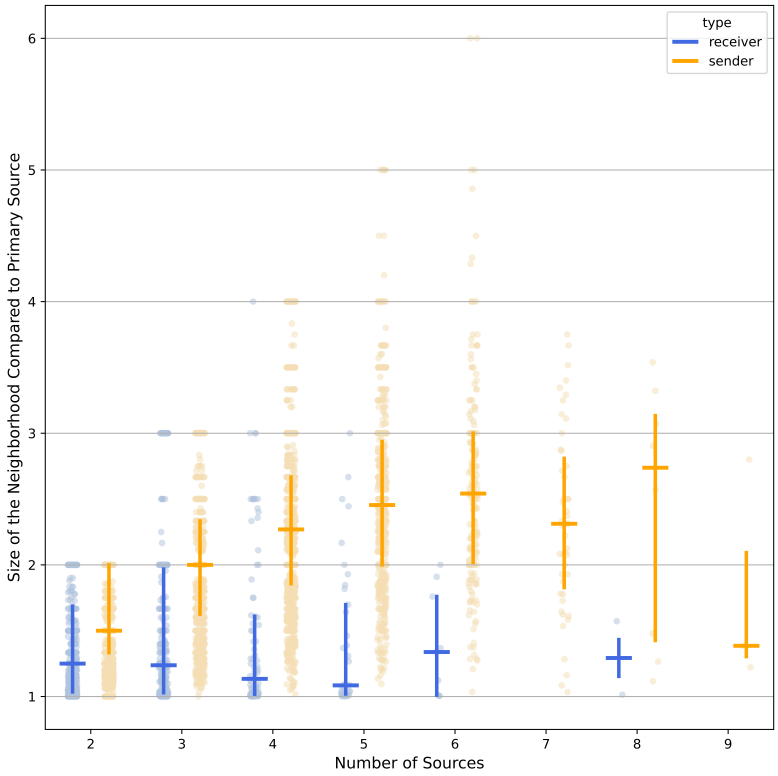

Fig. 4: The size of the sender and receiver neighborhoods of the actor and corresponding quartiles based on letters in all data sources of the actor compared to the primary data source with the largest number of neighbors.

Although the personal networks of actors can clearly become richer when combining multiple data sources, the number of shared sender-receiver pairs between data sources is relatively very low. This indicates that our understanding on the relationship between two people based on letters rarely changes when data sources are combined.

6 Data Quality Issues in Aggregated Collection KG

Completeness of the Data The completeness of the chosen metadata fields for the Person type actors increases as the number of letters an actor is associated with increases (see Fig. 5). When we increase the minimum degree, the proportion of actors that have available metadata grows across every chosen metadata field. The actors who have more letters preserved and digitized are also more likely to be found in external databases, from which some of the metadata is brought to the LetterSampo Finland KG. Of all Person type actors, only

18% are linked to external sources (the top row in Fig. 5). Only the gender of the actor is well known, 91% of them have a known gender assigned to them. The gender was assumed automatically, except in cases where only family name was known, there was mistakes in the spelling or the name was not commonly used in Finland.

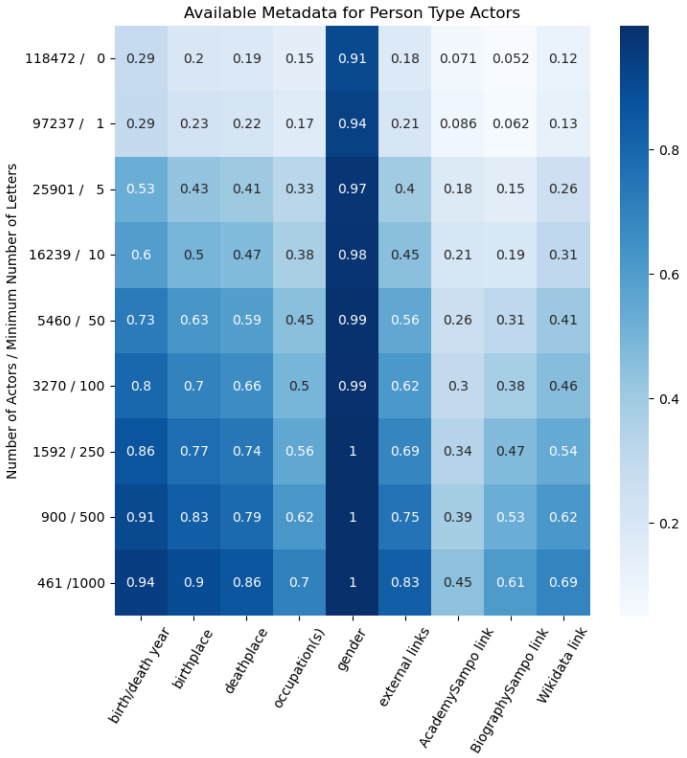

Fig. 5: Proportion of people who have the metadata and links to external sources available. Labels in y-axis tell the minimum number of letters actors are associated with and the number of such actors. Each cell tells the proportion of those actors that have the corresponding metadata available.

Fig. 6 shows the proportion of letters in each data source that have the sending date, language, and sending or target place of the letter available. The sending places of the letters are rarely known especially for the five largest data sources (the first five rows in Fig. 6), and the information about the places of receiving the letters is rare for all data sources except for the letters from the Postal Museum and Aalto University Archives, making it difficult to study the connections between places. The accuracy of places can vary from the residence of the actor to a country but is usually city, town or village.

Most letters have some date assigned to them, but as seen in Fig. 7, the accuracy of the date varies. The five largest data sources usually have a range of years or a year assigned as the sending date. The inaccuracy of sending dates hinders the temporal analyses, and if accurate sending dates are needed, one might want to focus on people like Elias Lönnrot, J. V. Snellman, Albert Edelfelt or Zachris Topelius whose designated letter collections usually have the exact sending dates available.

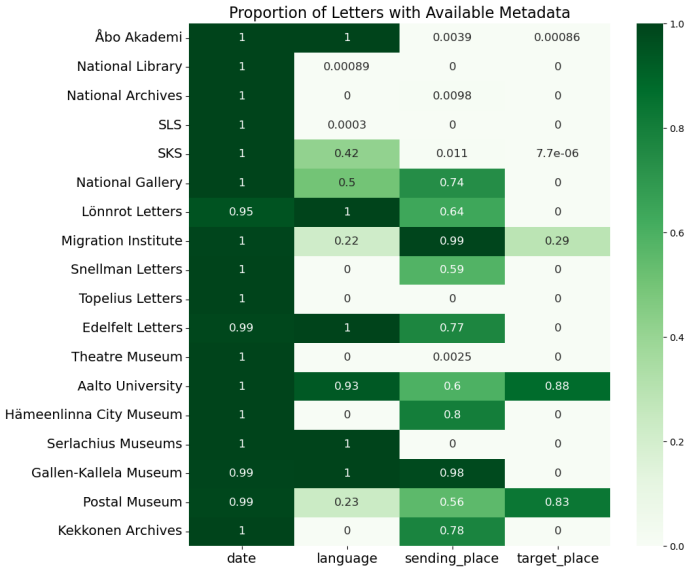

Fig. 6: Percentage of letters that have metadata available per source.

Detecting Duplicate Data and Misaligned Entities When combining letter metadata from multiple data sources, failing to link letters and actors that are present in multiple data sources to one entity leads to duplicate data if the link is missing, or misaligned entities if linking is erroneous. The biggest problem when aligning entities across or within the data source is the availability and quality of metadata for letters and people. If letters between two people are found in several sources but the contents, exact sending dates and places of the letters are not known, it is impossible to say for sure if the letters actually are the same. Similarly, if we know only the name of the actor that can vary due to name changes and different spellings, it can be difficult to link the actors if, for example, the name is common, the birth year of the actor is not known, or the collection does not give enough context.

There are more than 180 000 sender-receiver pairs in the data set. 1155 of them appear in two data sources, 18 pairs appear in three data sources, and only one pair in four data sources. Letter duplicates are hard to confirm due to varying

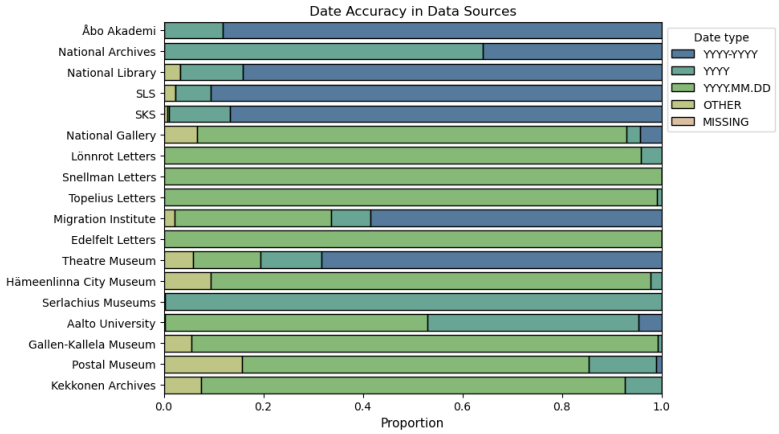

Fig. 7: Letter metadata availability and sending date accuracy per source.

sending data quality and missing sending places. Based on the exact sending dates and sending places, there is one possible duplicate letter between Edelfelt Letters and National Library, 4 within Lönnrot letters, and 22 between Lönnrot letters and Snellman letters. Based on only exact sending dates, there are 8 more possible duplicates between Lönnrot and Snellman. This can overestimate the letter-based link between Lönnrot and Snellman. Without taking into account the sending places, letters that have only the sending year available contain 2679 potential duplicate pairs.

As mentioned above, actors from different data sources are linked to *:ProvidedActor* class instances. During the linking process, the unclear cases were collected in a table. Each pair of actors was assigned a similarity score based on the actors' labels and other metadata, such as birth and death years and places, when applicable. The domain experts went through the table to determine if the actors are the same individual. In most cases, there was not enough information to link actors together with confidence, and these actor pairs are potential duplicates in the data. We expect that there are more potential duplicates than misaligned entities in the data.

Other cases of conflicting or erroneous data are letters whose sending date is before the birth or after the death of the sender and receiver of the letter. There are 7 letters that, according to the data, have been sent to the receiver years before their birth, and 88 letters that have arrived after the death of the receiver. About half of them have arrived a year or two after the death of the receiver of the letter, which might still be valid, but there are also clear errors where the sending date is tens of years after the death of the receiver of the letter. Similarly, there are 151 letters that have been sent outside the lifetime of the sender. These errors might occur during digitization or data harmonization.

7 Conclusions

Combining letter metadata from multiple data sources into one KG while harmonizing the data and aligning entities takes a lot of rigorous work. All historical data have parts of the data missing, and the process of combining data sources and related data transformations can give new problems such as duplicate data and misaligned entities, in this case especially among actors. Letters have fewer potential duplicates, as they tend to be present in only one data source. Lacking or missing metadata makes aligning entities and recognizing duplicates difficult.

By combining the data sources, we can have a more comprehensive view of the social networks of the time. Personal networks can grow manyfold in size for people who are present in many data sources, although for most people the growth of the neighborhood is modest. The growth is focused especially on outgoing links, that is, the people the actor sends the letters are added to the neighborhood. For people who are included in only one data source, combining multiple sources can help to clarify their social position among the whole social network. The links from actors to external sources bring in more information, allowing for more detailed analyses and helping to deduce why some actors have exchanged letters in the first place.

Based on our results, when one studies letters of actor based on one collection, one could expect that letters actor has sent are more likely missing or located in other sources than letters actor has received. In the case of LetterSampo Finland the inaccuracy of the sending dates and missing sending places make temporal or geographical analyses difficult, except for some individuals. Potentially missing letters and metadata have to be taken into account during analyses and interpreting results. Some of our results are visible on the LetterSampo Finland web portal where they can help researchers browsing the data to better understand it.

Here we focused on the metadata quality in the LetterSampo Finland and how letter collections enrich each other. Next steps include comparing results with other similar datasets, and moving from recognizing problems and errors in the data to automated or semi-automated methods for resolving them, when applicable. For other future work, the connections of the LetterSampo Finland KG to other external cultural heritage databases could be explored in more detail; for example, can some links be found between actors in external KGs that are not present in LetterSampo Finland KG and are there conflicting data between external data sources? The effects of combining multiple letter metadata sources on the whole network would be of interest.

Acknowledgments. Thanks to Jouni Tuominen, Ilona Pikkanen, and other co-workers in the CoCo project for creating the CoCo KG and fruitful discussions. This work was mainly funded by the CoCo project supported by the Research Council of Finland. Partial funding was received from the European Union – NextGenerationEU instrument under grant number P3C3I6 for the national FIN-CLARIAH/DARIAH-FI initiative.

Bibliography

[1] Behkamal, B., Kahani, M., Bagheri, E., Jeremic, Z.: A metrics-driven approach for quality assessment of linked open data. Journal of Theoretical and Applied Electronic Commerce Research **9**(2), 64–79 (2014). https://doi.org/10.4067/S0718-18762014000200006

[2] Drobac, S., Enqvist, J., Leskinen, P., Wahjoe, M.F., Rantala, H., Koho, M., Pikkanen, I., Jauhiainen, I., Tuominen, J., Paloposki, H.L., Mela, M.L., Hyvönen, E.: The laborious cleaning: Acquiring and transforming 19th-century epistolary metadata. In: Digital Humanities in the Nordic and Baltic Countries Publication, DHNB2023 Conference Proceeding. vol. 5, pp. 248–262. University of Oslo Library, Norway (2023), `https://doi.org/10.5617/dhnbpub.10669`

[3] Düring, M.: How Reliable are Centrality Measures for Data Collected from Fragmentary and Heterogeneous Historical Sources? A Case Study. In: The Connected Past: Challenges to Network Studies in Archaeology and History. Oxford University Press (03 2016). https://doi.org/10.1093/9780198748519.003.0011

[4] Edwards, G., Crossley, N.: Measures and meanings: Exploring the ego-net of Helen Kirkpatrick Watts, militant suffragette. Methodological Innovations Online **4**(1), 37–61 (2009). https://doi.org/10.1177/205979910900400104

[5] Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) The Semantic Web: Research and Applications. pp. 87–102. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[6] 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

[7] Hyvönen, E., Leskinen, P., Tuominen, J.: LetterSampo – historical letters on the semantic web: A framework and its application to publishing and using epistolary data. Journal on Computing and Cultural Heritage **16**(1) (2023). https://doi.org/10.1145/3569372

[8] Hyvönen, E., Leskinen, P., Poikkimäki, H., Rantala, H., Leal, R., Tuominen, J., Drobac, S., Koho, O., Pikkanen, I., Paloposki, H.L.: Searching, exploring, and analyzing historical letters and the underlying networks:LetterSampo Finland — Finnish 19th-century letters on the semantic web. In: Digital Humanities in Nordic and Baltic Countries 2025 (DHNB 2025), Postproceedings. University of Oslo Library, Norway (2025), in press

[9] Hyvönen, E., Tuominen, J.: 8-star linked open data model: Extending the 5-star model for better reuse, quality, and trust of data. In: Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024). vol. 3759. CEUR Workshop Proceedings (September 2024), `https://ceur-ws.org/Vol-3759/paper4.pdf`

[10] 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
[11] 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
[12] Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web **11**(2), 255–335 (2020). https://doi.org/10.3233/SW-180333
[13] van Miert, D.: What was the Republic of Letters? A brief introduction to a long history (1417–2008). Groniek **204/205**, 269–287 (11 2016)
[14] Mountantonakis, M., Tzitzikas, Y.: High performance methods for linked open data connectivity analytics. Information **9**(6), 134 (2018). https://doi.org/10.3390/info9060134
[15] Nayak, A., Božić, B., Longo, L.: Linked data quality assessment: A survey. In: Xu, C., Xia, Y., Zhang, Y., Zhang, L.J. (eds.) Web Services – ICWS 2021. pp. 63–76. Springer International Publishing, Cham (2022)
[16] Poikkimäki, H., Leskinen, P., Hyvönen, E.: Exploring cultural heritage knowledge graphs – case of correspondence networks in the Grand Duchy of Finland 1809–1917. Digital Humanities in the Nordic and Baltic Countries Publications **7**(2) (Mar 2025). https://doi.org/10.5617/dhnbpub.12289
[17] Riehle, A., Preiser-Kapeller, J.: Letters and Network Analysis, vol. 7. BRILL, Leiden; Boston : (2020)
[18] Ryan, Y.C., Ahnert, S.E.: The Measure of the Archive: The Robustness of Network Analysis in Early Modern Correspondence. Journal of Cultural Analytics **6**(3) (7 2021). https://doi.org/10.22148/001c.25943
[19] Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (May 1997). https://doi.org/10.1145/253769.253804
[20] Sugimoto, G.: Instance level analysis on linked open data connectivity for cultural heritage entity linking and data integration. Semantic Web **14**(1), 55–100 (2022). https://doi.org/10.3233/SW-223026
[21] 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
[22] Yu, Y., Romero, D.M.: Does the use of unusual combinations of datasets contribute to greater scientific impact? Proceedings of the National Academy of Sciences **121**(41) (2024). https://doi.org/10.1073/pnas.2402802121
[23] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey: A systematic literature review and conceptual framework. Semantic Web **7**(1), 63–93 (2015). https://doi.org/10.3233/SW-150175