Casting the net far and wide: Aggregating and harmonizing epistolary metadata in collaboration with cultural heritage institutions

Drobac, Senka
senka.drobac@aalto.fi
Aalto University (Semantic Computing Research Group (SeCo)), Finland

Enqvist, Johanna
johanna.enqvist@finlit.fi
The Finnish Literature Society, Finland; University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland

Leskinen, Petri
petri.leskinen@aalto.fi
University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland; Aalto University (Semantic Computing Research Group (SeCo)), Finland

Wahjoe, Muhammad Faiz
muhammad.wahjoe@aalto.fi
Aalto University (Semantic Computing Research Group (SeCo)), Finland

Rantala, Heikki
heikki.rantala@aalto.fi
Aalto University (Semantic Computing Research Group (SeCo)), Finland

Koho, Mikko
mikko.koho@aalto.fi
Aalto University (Semantic Computing Research Group (SeCo)), Finland

Pikkanen, Ilona
ilona.pikkanen@finlit.fi
The Finnish Literature Society, Finland

Jauhiainen, Iida
iida.a.jauhiainen@helsinki.fi
The Finnish Literature Society, Finland

Tuominen, Jouni
jouni.tuominen@aalto.fi
University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland; Aalto University (Semantic Computing Research Group (SeCo)), Finland

Paloposki, Hanna-Leena
hanna-leena.paloposki@kansallisgalleria.fi
The Finnish Literature Society, Finland; Finnish National Gallery, Finland

La Mela, Matti
matti.lamela@helsinki.fi
University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland; Uppsala University, Sweden

Hyvönen, Eero
eero.hyvonen@aalto.fi
Aalto University (Semantic Computing Research Group (SeCo)), Finland; University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland

This paper describes the process of gathering, aggregating, harmonizing, and publishing epistolary metadata through collaboration with Finnish cultural heritage (CH) organizations in order to create an inclusive archive for bottom-up analyses of 19th-century epistolary culture in the Grand Duchy of Finland (1808/09-1917). The authors are working in the digital humanities consortium project Constellations of Correspondence (CoCo) (Tuominen et al. 2022). The unified metadata collections are harmonized, linked, enriched, and published on a Linked Open Data (LOD) service, and as a semantic web portal.

In Europe, there are several digital humanities projects using well-curated metadata (detailed information about senders, recipients, dates, and places) from edited letter collections - like CKCC (van den Heuvel 2015), correspSearch (Dumont 2016, Dumont et al 2021), the Early Modern Letters Online (EMLO) (EMLO, 2022, Hotson / Wallnig 2019), Norkorr (Rockenberger 2019), and SKILLNET (SKILLNET, 2023). In our project, most of the data come from unpublished collections scattered around different Finnish CH organizations. Collaboration with these CH organizations is pivotal for the successful outcome of the project. It requires a dialogue with them throughout the whole project period in the form of seminars and site visits, as well as sharing blogs and newsletters, also after the organizations have provided their letter metadata, as illustrated in Figure 1. We have also already seen that some of the participating organizations are prepared to clean their metadata or catalogue previously uncatalogued archival material to provide better and more metadata for the project. We will discuss this two-way process using the Finnish National Gallery as a case study. An important challenge yet to be studied profoundly is, if and how the CoCo project will be able to deliver to the CH organizations their metadata in an enriched format.

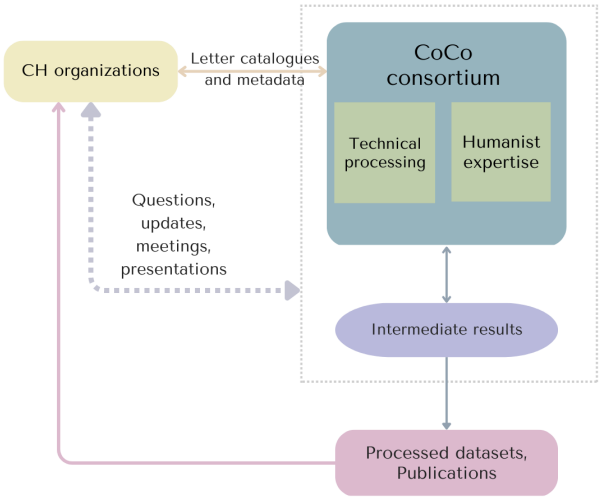

Figure 1. A diagram illustrating the dialogue between the consortium and the CH orgnaizations.

In the first phase of the project, we conducted a survey that was sent to over 100 CH organizations (extending from small local museums to official central archives). The paper describes how the information was collected and how the survey was constructed in order to provide us with detailed enough information regarding their 19th-century collections and metadata formats. At the same time, we had to keep the query succinct in order to make the answering as effortless as possible. Some of the results are shown in Figure 2.

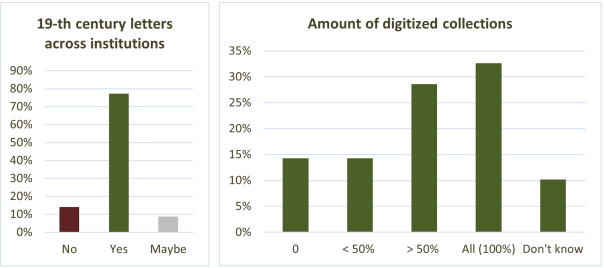

Figure 2. According to the survey findings, 77% of institutions have letter collections from the 19th century (shown in the left graph). Out of these collections, as indicated in the right graph, 14% remain entirely non-digitized, 14% have less than 50% digitized, 28% have digitized over 50%, and 32% have fully digitized their collections. Roughly 10% of the institutions are unsure about the digitization status of their collections.

As to the data processing, we began with more than 350 000 letters, from eight different sources, each in its own digital format. Although the received data is mostly structured, we needed to parse running text to retrieve metadata in nearly every collection. Moreover, we had to analyze each dataset and identify possible structural mistakes. Furthermore, some records required Natural Language Processing to get actor names (e.g. senders, recipients) in dictionary format. The most difficult task has been to process word files which contain correspondence metadata in a variety of formats, easily understandable to humans but difficult for computational processing.

A harmonizing data model for epistolary metadata collections was developed, which builds on international standards like CIDOC CRM to promote interoperability. The most central classes are Letter, Place and Actor. Also, provenance and archival information are included.

Finally, the actor data is enriched by linking it to external databases like Wikidata and the Finnish AcademySampo and BiographySampo. These external sources provide detailed biographical information, e.g., times and places of birth and death, name variations, occupations, or genealogical relationships. Information present in the letter metadata like actor names and times of sending and receiving is used for matching entities between our data and the external databases, and further to reconcile the actors between data sources.

Bibliography

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

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

EMLO (2022): URL: http://emlo.bodleian.ox.ac.uk.

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

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

Rockenberger A. et al. (2019): Norwegian correspondences and linked open data, in: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, volume 2364 of CEUR Workshop Proceedings, pp. 365–375. URL: http://ceur-ws.org/Vol-2364/33_paper.pdf.

SKILLNET (2023): Sharing Knowledge in Learned and Literary Networks – The Republic of Letters as a Pan-European Knowledge Society, 2022. URL: https://skillnet.nl.

Tuominen J. et al. (2022): 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: 6th Digital Humanities in Nordic and Baltic Countries Conference. URL: http://ceur-ws.org/Vol-3232/paper41.pdf.