The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata

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

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

Abstract

This paper describes the process of gathering, aggregating, harmonizing, and publishing epistolary metadata from 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)* project [1] that aggregates and publishes 19th-century epistolary metadata from scattered collections of Finnish CH organizations. The unified collections are harmonized, linked, enriched, and published on a Linked Open Data (LOD) service, and as a semantic web portal.

Although this project is dealing with Finnish epistolary metadata (or metadata that has ended up in the Finnish archives and museums), we believe that our experiences have wider significance. In Europe, there are several digital humanities projects harvesting well-curated metadata (detailed information about senders, recipients, dates, and places) from edited letter collections - like correspSearch and Norrkor, and in some cases, the aim is to reach out to letter catalogues of CH organizations.

On the more general level, the paper participates in the ongoing discussion regarding the initial phases of data-intensive research and how this time-consuming "data work" should be described, understood, and credited. As Ahnert et al. [2] have recently argued, *"the lack of discussion around such practices ... has increased mistrust of quantitative approaches in the arts*

The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023)
senka.drobac@aalto.fi (S. Drobac)
0000-0002-7645-3079 (S. Drobac); 0000-0003-0901-7987 (J. Enqvist); 0000-0003-2327-6942 (P. Leskinen); 0000-0002-4716-6564 (H. Rantala); 0000-0002-7373-9338 (M. Koho); 0000-0001-9435-7163 (I. Pikkanen); 0000-0003-4789-5676 (J. Tuominen); 0000-0003-1412-8622 (H. Paloposki); 0000-0003-0340-9269 (M. La Mela); 0000-0003-1695-5840 (E. Hyvönen)

and humanities. The only way to deal with this is to begin talking about the labour of cleaning and to communicate its significance as an intellectual contribution."

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.

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 400 Word files provided by the National Library of Finland, 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.

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] J. Tuominen, M. Koho, I. Pikkanen, S. Drobac, J. Enqvist, E. Hyvönen, M. La Mela, P. Leskinen, H.-L. Paloposki, H. Rantala, 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, short paper., 2022. URL: http://ceur-ws.org/Vol-3232/paper41.pdf.

[2] R. Ahnert, S. E. Ahnert, C. N. Coleman, S. B. Weingart, The Network Turn: Changing Perspectives in the Humanities, Cambridge University Press, 2020.