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], 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*

[5] *Uppsala University, Sweden*

Abstract

The paper documents the process of collecting, consolidating, and publishing epistolary metadata from Finnish cultural heritage organizations to create an archive for bottom-up analyses of 19th-century epistolary culture. We describe and discuss the data survey that was conducted to gather information about available letter collections across Finland, as well as the cleaning and harmonizing of over 350,000 letters from twelve different sources in various digital formats. We have also developed a data model that combines event-based and letter-based aspects of the metadata. Furthermore, the paper contributes to the ongoing discussion of the initial phases of data-intensive research and the importance of discussing the labor of cleaning data. We believe that our experiences described in this paper can have wider significance for other digital humanities projects in Europe.

Keywords

Letter Metadata, Semantic web, Linked Open Data, 19-th Century

1. Introduction

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)* [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.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 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)

CEUR Workshop Proceedings (CEUR-WS.org)

In what follows we will scrutinize the different phases of the data acquisition and processing. First, we will discuss a data survey that was sent to a wide variety of Finnish CH organizations in order to acquire systematic and comparable information as to their collections. Second, we will describe the stages of processing and cleaning the epistolary metadata. We began with more than 350 000 letters, from twelve 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.

Furthermore, we explain the efforts made to create a harmonizing data model for epistolary metadata collections that adhere to international standards. The data model is designed to support modeling of the relevant properties of letter metadata collected from source datasets, to promote interoperability, and to support efficient use of data in e.g. SPARQL queries and the semantic portal developed during the project.

We will round off the paper by discussing the initial phases of data-intensive research and how this time-consuming "data work" should be described, understood, and credited. [2]

Although this paper describes and discusses the processing of Finnish epistolary metadata (or metadata that has ended up in the Finnish archives and museums), we believe that our experiences may have wider significance. In Europe, there are several digital humanities projects that harvest well-curated metadata (detailed information about senders, recipients, dates, and places) from edited letter collections – like Europeana[1] [3], Kalliope Catalogue[2], The Catalogus Epistularum Neerlandicarum[3], Electronic Enlightenment[4], ePistolarium[5] [4], SKILLNET[6], correspSearch[7], the Mapping the Republic of Letters project[8], NorKorr - Norwegian Correspondences and Linked Open Data [5], and Early Modern Letters Online (EMLO)[9] [6, 7, 8]. Bruneau et al. discuss applying Semantic Web Technologies to modelling the correspondences of French scientist Henri Poincaré and publishing on an online portal[10] [9]. The data work described in this article can therefore serve as a precedent for future projects, that set out to acquire letter metadata from the collections of cultural heritage organisations on a wider scale.

[1] http://www.europeana.eu

[2] http://kalliope.staatsbibliothek-berlin.de

[3] http://picarta.pica.nl/DB=3.23/

[4] http://www.e-enlightenment.com

[5] http://ckcc.huygens.knaw.nl/epistolarium/

[6] https://skillnet.nl

[7] https://correspsearch.net

[8] http://republicofletters.stanford.edu

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

[10] http://henripoincare.fr/s/correspondance/page/accueil

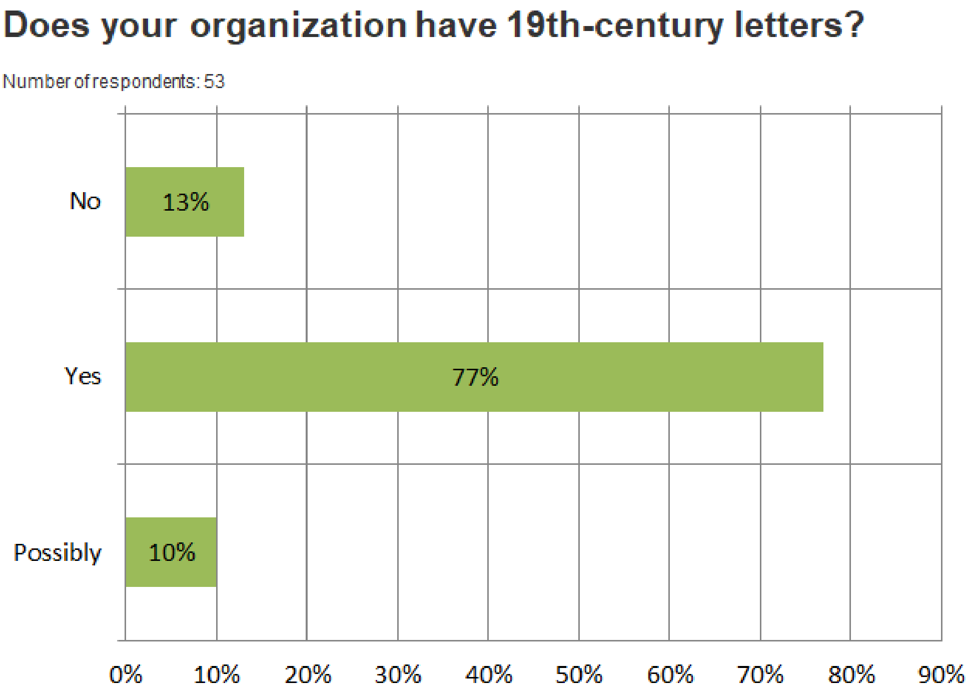


Figure 1: The Webropol survey for Finnish CH organizations: percentage of organizations governing 19th-century letters

2. Data Acquisition

In the first phase of the project, when planning the gathering of metadata from the Finnish CH organizations, we soon realized that to form an overview of the field we must try to investigate how much 19th-century correspondence there actually is in the Finnish CH organizations – archives, libraries and museums. This information was not in general pre-available from the organizations themselves; we will get to the reasons later in this article.

To get the needed information and to approach the possible governors of 19th-century letters, we composed a Webropol metadata survey with 18 questions, including both fundamental questions about the existing letter material and more detailed questions regarding the letters and their metadata in individual CH organizations. The survey stresses that in order to obtain a reliable overall picture, we would also welcome responses from those organisations that do not have 19th century letter collections.

Since the early 2022, we have sent links to the survey, and have partially re-sent them to 102 CH organizations, ranging from large institutions to small local museums and archives. By the end of February 2023, we had received 53 responses to the survey. We consider this response rate to be significant as it appears that the majority of CH organizations with large letter collections have responded. There is also a public web link available, which has been shared via blogs and

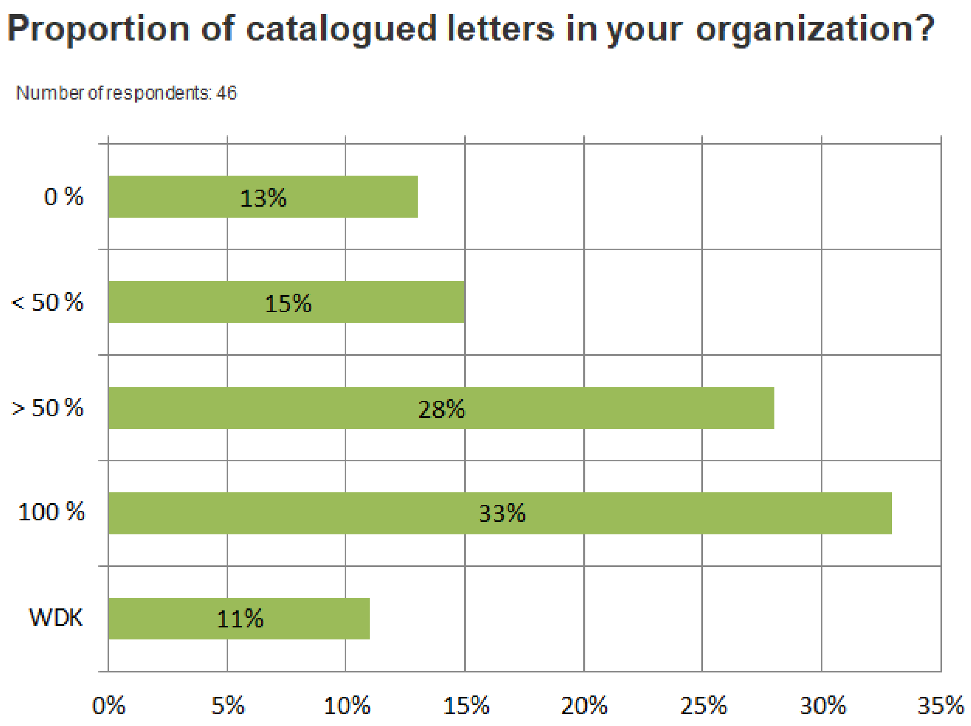


Figure 2: The Webropol survey for Finnish CH organizations: proportion of catalogued 19th-century letters

in social media. The biggest national CH organizations, the National Archives of Finland, the National Library of Finland and the Finnish National Gallery, had already in advance agreed on submitting their letter metadata and did not participate in the survey.

Once we analyzed the results of the survey, we found that 77% of respondents indicated that they have 19th-century letters in their collections, as illustrated in Figure 1. Furthermore, the majority of these organizations expressed their interest in cooperating by submitting their letter metadata. As of writing this article, we have received metadata from 12 organizations. More material is coming in and the data migration has already been agreed with five more organizations. Gathering metadata is thus an ongoing process and will continue in 2023–24.

While trying to get the idea of the overall number of preserved 19th-century letters, we have come up to the fact the CH organizations very often do not know the exact number of letters in their collections. It is maybe characteristic that 10% of the respondents answered that they possibly have 19th-century letters. The answers to the question of the percentage of catalogued letters in their collections give an explanation to this. As Figure 2 illustrates, only 33% of the respondents with these collections have catalogued all their letter material, 28% over a half of them, 13% not at all and 11% could not give estimation. This tells about the everyday realities in CH organizations: they do not have enough resources to organize and catalogue all their

Table 1
Overview of the collections being processed, including the collection name, total number of letters, and format of the data received. The first three collections are well-curated and edited, while the remaining collections are obtained from various institutions.

Name	Size (Letters)	Format
Albert Edelfelt	1 600	JSON Web API
Elias Lönnrot	6 247	JSON Web API
J.V. Snellman	1 514	RDF
Svenska litteratursällskapet i Finland (SLS)	43 000	XLS files
National Gallery	9 976	CSV
Finnish Art Society	1 147	XLSX
National Archive	295 000	CSV
Åbo Akademi	366 614	XML
The Finnish Literature Society (SKS)	37 676	XLSX
HS Foundation	2 500	CSV
Postal Museum	50	XLSX
National Library	unknown	Word files

archival collections.

Computational methods can naturally only make use of metadata in a structured form, i.e. in a form that can be processed computationally. In addition to the lack of overall quantitative data on the collections, we did not know beforehand how much of the data has been described (catalogued) by the archivists and what proportion of this metadata is computationally accessible. According to our survey, only 35% of the existing catalogues on 19th century letters are in some electronic form: in databases, as word or excel documents.

The answers to the survey reflect the fact that the letter collections have accumulated in archival collections over decades, sometimes centuries. Consequently the metadata production has deep temporal layers too. Even within a single collection, correspondence may have been described at different stages and on different platforms. Descriptive practices and concepts have evolved over time, but also in relation to the specificity of the material being described and the creative solutions of individual archivists. Today's digital systems are promoting the production of more structured descriptive information. However, it is not self-evident that the metadata fields of the different archival systems are comparable or that the subsequent extraction of data has been planned and tested when organisational databases and archival information systems have been built.

Overall, the varying practices of storing epistolary metadata challenge the computational use of it. In the next section we describe the various solutions of processing the data, depending on the format and on the varying contents and structures of the letter catalogues.

3. Data Cleaning and Transformation

An overview of the collections that have been received and are currently being processed is presented in Table 1. The table has three columns: the first column lists the names of the collections, the second column shows the total number of letters in each collection, and the last column specifies the format of the received data. The first three rows, which are highlighted in a light gray color, correspond to well-curated published edited collections. The remaining collections are obtained directly from CH institutions, some as data dumps from their internal systems and others as copies of their storage records.

To harmonize all these different datasets, we have created a fully automatic transformation pipeline, like illustrated in Figure 3. The pipeline comprises several stages, starting with the processing of each received dataset into an intermediary RDF (Resource Description Framework) format, which contains literal values. The next step involves harmonizing the data with the CoCo Data Model, which we have developed based on international standards. To enrich the data, we link the recognized actors and places with external resources. At this point, it is possible and desirable to deduplicate actor and place names, a process that we plan to develop in the future. Finally, the transformation pipeline produces a harmonized dataset of correspondence metadata, which is structured and optimized for accurate and efficient analysis.

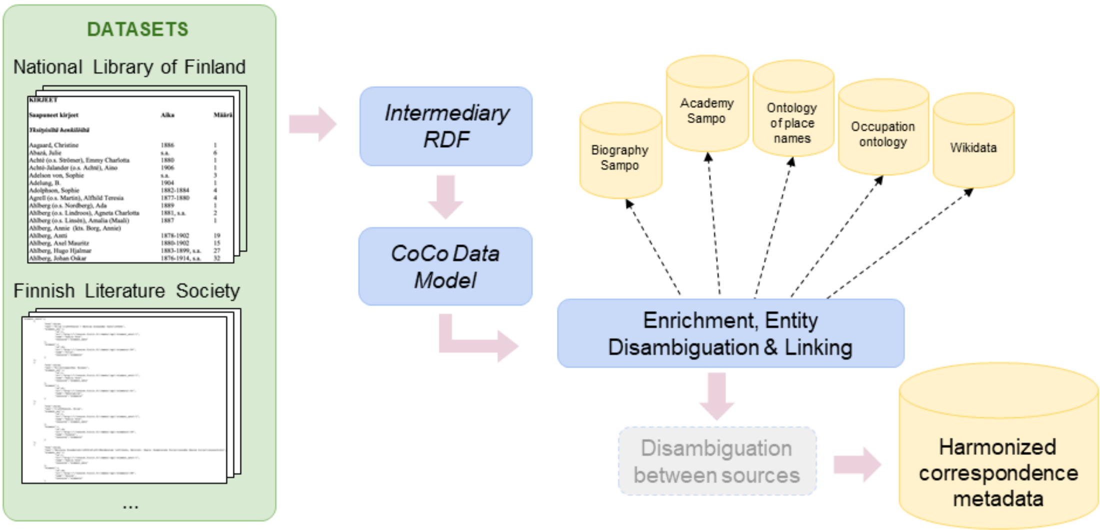

DATASETS
National Library of Finland
Finnish Literature Society
...
Intermediary RDF
CoCo Data Model
Biography Sampo
Academy Sampo
Ontology of place names
Occupation ontology
Wikidata
Enrichment, Entity Disambiguation & Linking
Disambiguation between sources
Harmonized correspondence metadata

Figure 3: Illustration of the fully automatic transformation pipeline used to harmonize heterogeneous datasets. The pipeline involves several stages, including conversion into an intermediary RDF format, harmonization with the CoCo Data Model, linking with external resources, and deduplication/disambiguation (planned for future development). The resulting harmonized dataset of correspondence metadata is structured and optimized for analysis.

Cleaning and harmonizing datasets

In the first step, we created a separate transformation process for every dataset where we locate and extract relevant information. Given the substantial variations in the nature and details of the correspondences across the datasets, we analyzed the first 11 datasets to determine a comprehensive list of extractable properties. The resulting list is presented in Table 2.

Table 2
A list of the properties we have collected from the source datasets. For easier comprehension, we can divide properties in three groups: information about sender and recipient, correspondence information, and archival information. The properties we searched for and extracted from the datasets are listed in the right column. (Not all properties are always present.)

Group of information	Collected properties
Sender and recipient	Full name, first name, last name, particle of nobility, date of birth, date of death, gender, occupation, type (e.g. person, company, family)
Correspondence information	Date of sending, amount of letters, place of sending, place of receiving, language of the letter, letter type (e.g. letter, postcard, telegram), content of the letter, translation of the letter, person reference (i.e. people mentioned in the letter), place reference (i.e. mentioned places)
Archival information	Record id, archival fonds, series/letter collection

In almost all the cases, the source datasets have information about both sender and recipient (i.e. actors). The most important thing for us is to get the full name, which can be written in different formats (e.g. *"Walleen, Carl Johan"*, *"Carolina Carlstedt"*, *"C. M. Creutz"*). Sometimes, the actor is a family (e.g. *"perhe Carlstedt"*), or institution (e.g. *"Königliche Akademie der Künste zu Berlin"*). Occasionally, there is also information about a person acting on behalf of an institution (e.g. *"Snellman, Johan Vilhelm / STY:n taloustoimikunta"*). Besides the name, additional information about the actors (such as gender and date of birth and death) is available only in few datasets, but we try to acquire it whenever available because it is important for the later disambiguation of the actors.

The available information on correspondence varies considerably across datasets. Temporal information is generally available, either in the form of a full date or just a year of sending. In most cases, however, the number of letters is reported as the total number of letters sent within certain boundary years. Occasionally, information on the place of sending is available, but information on the destination is rare. Language and letter type information is also available in some datasets, ranging from regular letters, postcards, telegrams, to less common letter types such as invitations. Information-rich datasets may even contain summaries or full contents of letters, as well as information on persons and places mentioned within them.

The archival information that we aim to capture and present includes the correspondence ID, if available, or another identifier that can pinpoint the precise location in the received dataset, such as the row ID for CSV files. We also retain the name of the archive, such as *"Elisabeth Järnefeltin arkisto"* and the name of the series or collection of letters, such as *"COLL.101.15."*.

To maintain provenance, we also preserve information on the actors as they are originally recorded, as well as the subset of the original dataset, such as the entire row in a CSV table. This approach ensures that any mistakes made during automatic processing can be detected and that the final user can view the original records. It enhances the trustworthiness of the harmonized dataset and allows for the easy reproduction of the process. In other words, our fully automatic process is transparent, and anyone can see the operations being performed. In case of the word files, we document the manual processing and once we have processed the data, we will publish a document listing all the changes that we have made. This part is the least transparent because it is challenging to show the original segments, but we hope that the manual will provide the user with a clear understanding of the cleaning that has been done.

We have an issue in dealing with families and institutions as actors in the data. For instance, some datasets only specify the recipient as "family Carlstedt", without providing details on the specific person within the family who received the communication. Letters could be written to the whole family or a group of family members and as we do not know if there were possible individual letter receivers in these cases, we have to retain the recipient as recorded. Likewise, when institutions are involved, we often lack information on the particular sender or recipient of a letter.

Structural issues

We have received certain datasets as data dumps from internal institutional database systems. However, the dumping process occasionally encounters errors, as evidenced by error reports found in some datasets, such as:

```
  No signature of method: com.zetcom.mp.service.provider.data.search.impl.
lucene.domainFacade.ControlledVocabularyNodeImpl.plus()  is  applicable
for argument types: (java.lang.String) values: [, ] Possible solutions:
is(java.lang.Object), split(groovy.lang.C
```

In addition, we have observed that some cells have been shifted left or right in a couple of CSV datasets. It is crucial to detect such errors and implement appropriate handling mechanisms to avoid obtaining anomalous results.

Names and persons without names

In some cases, we need to extract an actor's name from free text, as exemplified by phrases such as "Elisabeth Järnefelt's fonds", "Ensio Hiitonen's conversations" or "Aarno Durchman's letters to Sigrid Duchman" (originals in Finnish or Swedish). While obtaining the basic form of the name from the genitive case is relatively simple in English, it is much more complex in Finnish. In Finnish, words not only receive a suffix, but also undergo changes. For instance, here are several examples of genitive cases and basic forms in Finnish: "Järnefeltin – Järnefelt", "Suolahden – Suolahti", "Hiitosen – Hiitonen ".

To obtain the base form of names, we used the FinnPos lemmatizer [10]. The lemmatizer uses Finnish morphology Omorfi [11] but can also lemmatize unknown words, making it an excellent tool for handling names, particularly given the presence of foreign names in our data sets. Nonetheless, the lemmatizer may occasionally make mistakes, which is why we manually corrected the results and stored them in a dictionary for easy reuse.

Moreover, in all wide correspondences, there are also unknown senders, persons who have not been be identified by close-reading the letters or from other correspondence. Currently, we have identified approximately 30 different ways to label unknown individuals in datasets, with the term "Tuntematon" being the most frequent one. Having unidentified senders may be

Table 3
Example from the HS foundation CSV file. We are showing three rows and two columns, which contain details about senders and their sent letters. In the first row, an asterisk (*) indicates a telegram (sähke), while the letter S (S = saapuneet kirjeet) signifies that the listed letters are received ones. The recipient is the fonds' holder and that column is not displayed in this excerpt.

A cell in CSV file containing correspondence information	Other info
Nekton, Toivo H. (Broolyn, New York): S: 16.9.*, 20.9.* ja 13.10.1905	* sähke
Kirjeenvaihto Manda ja Juho Soinin kanssa: S: 25.7.1897.	S = saapuneet kirjeet
Kirjeenvaihto Edla Soldanin kanssa: S: 23.5.1870, 12.8., 4.9., 25.9., 21.10. ja 30.11.1872, 1.2., 7.3., 18.3., 1.4., 7.5., 28.5./16.6., 6.9., 9.10., 11.10. ja 1.11. 1873, Heikin päivänä, 1.3., 18.3., 2.4., 12.4., 2.5., 4.7., 23.7., 2.9., 13.9., 28.9., 7.10. ja 22.10.1874, 3.2., 4.4., 6.5., 24.7., 21.11.1875, 6.2., 26.3., 22.10.1876, 29.4., 1.7., 25.10., 30.12.1877, 22.9. ja 29.9. 1878, 1.4. ja 20.9.1879, 13.3.1880, 3.1. ja 5.5.1884. (5 nippua) Kirjeistä on myös valokopiot. (1 nippu)	S = saapuneet kirjeet

frustrating for the user of a single collection, but it becomes a problem when we bring together various fonds with numerous unknown senders. This means that each unidentified person must be given a unique and identifying "unknown identifier".

Parsing free text

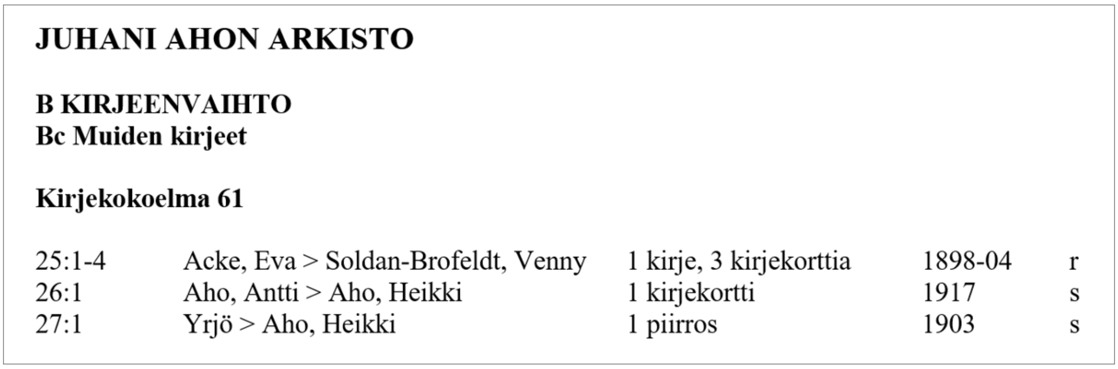
Parsing free text can be a challenge in certain datasets, even if they appear to be structured, such as in CSV format. While some examples, such as the National Gallery files, may contain sender and receiver information in a formatted text format (e.g., "Ingrid Carlstedt, sender; Mikko Carlstedt, recipient"), more complex cases may involve actors, places, and individual letter dates, with varying formats across cells. An example of such complexity is shown in Table 3. The first row indicates that the sender was "Nekton, Toivo H.", the place of sending was "Brooklyn, New York", and two telegrams (*sähke*) were sent in September 1905 along with one letter in October 1905. The second row involved two senders, "Manda and Juho Soini", who sent one letter. The third row shows that "Edla Soldan" sent multiple letters on different dates. The recipients of the letters are not displayed in this table, as that information was straightforward to process in a separate column.

The letter collection by the Finnish Art Society contained brief summaries of the letter content often mentioning related people, organizations or places. These summaries were written in a free text format, like for example "Walter Runeberg's scholarship to travel to Rome" or "Altarpiece for church in Kalajoki, Adolf von Becker". In processing NLP tools [12] based on FinBERT [13] were used to extract these references to named entities.

Processing information provided in this manner can be difficult, especially when dealing with large and varied datasets. To successfully parse the data, we must develop parsers to recognize different patterns, while also employing a lemmatizer to obtain the basic form of a name and using name processing algorithms to ensure consistency in the format of names.

Word files

Some prominent organizations like the National Library of Finland (NL) and the Swedish Literary Society in Finland (SLS) still maintain traditional word-format catalogues as the ”user interface” to their epistolary collections (in the case of NL, the letter metadata only exists in Word format). Such catalogues are perfectly suited to the needs of human users. They are sufficiently consistent to allow information seekers to quickly, on the basis of previous knowledge, to grasp their logic. However, they provide automatic, algorithmic reading with a set of specific challenges.



JUHANI AHON ARKISTO

B KIRJEENVAIHTO
Bc Muiden kirjeet

Kirjekokoelma 61

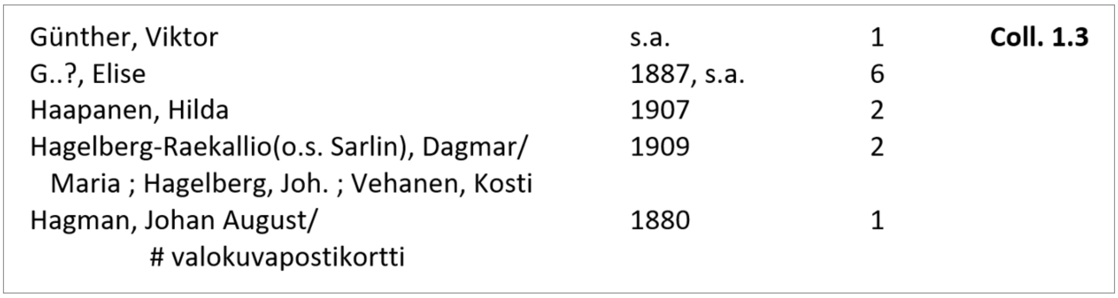
25:1-4	Acke, Eva > Soldan-Brofeldt, Venny	1 kirje, 3 kirjekorttia	1898-04	r
26:1	Aho, Antti > Aho, Heikki	1 kirjekortti	1917	s
27:1	Yrjö > Aho, Heikki	1 piirros	1903	s

Figure 4: Juhani Aho’s correspondence catalogue contains information about letter exchange (Kirjeenvaihto) between other people (Muiden kirjeet). Additionally, this particular catalogue includes information about the type of documents exchanged, such as letters (kirje), postcards (kirjekortti(a)), and drawings (piiros). The fonds of Aho is kept at the Finnish Literature Society, Helsinki.

The catalogues have been created by different archivists in a wide variety of organizations with their own cataloguing practices over the decades, resulting in inconsistent formatting. Typically these files begin with the name and brief biography of the records creator, the primary person whose archive it is, followed by information about the various documents in their archive. The catalogue section that is of particular interest to us is the Letter Exchange, which is usually categorized into subcategories such as Received Letters, Sent Letters, Letter Concepts (sometimes “Unsent Letters”), and Letter Exchange between “Other Individuals”. In other words, the archive (and the correspondence catalogue) of the prominent 19th-century author Juhani Aho, held in the archive of the Finnish Literature Society, contains letters exchanged by his wife Venny Soldan-Brofeldt and her acquitance Eva Acke, as shown in Figure 4. Except for the last subcategory, all the other subcategories usually contain four columns separated by tabs or spaces, with details about the name of an actor, time, quantity of letters, and the signum (the collection’s reference code). The section containing letter exchange between other individuals includes details about two actors, the sender and the recipient, along with the time, quantity, and the signum.

Parsing the files automatically poses several challenges due to various issues in the files. Firstly, the general document structure is diverse. Some catalogues contain archival information not only on one person but also on their spouse and other family members. Additionally, not all documents follow the same structure, and catalogues with abundant material often have a wide variety of formatting. The subsections of the letter exchange also do not have naming conventions, but require a deeper semantic language understanding due to their creativity.

Inconsistencies at the line level further compound the parsing challenge. For example, sometimes information about one letter correspondence spans multiple lines, either due to insufficient space or additional information such as comments or place of sending. In some cases, there is even information on some other correspondence between other people.



Günther, Viktor	s.a.	1	**Coll. 1.3**
G..?, Elise	1887, s.a.	6	
Haapanen, Hilda	1907	2	
Hagelberg-Raekallio(o.s. Sarlin), Dagmar/	1909	2	
Maria ; Hagelberg, Joh. ; Vehanen, Kosti			
Hagman, Johan August/	1880	1	
# valokuvapostikortti			

Figure 5: A sample of Ida Aalberg's collection of received letters from the National Library Word files, which has undergone manual editing.

Since the data is highly heterogeneous, a complex parser is required to process it. However, creating such a parser would be time-consuming and labor-intensive. Therefore, we have opted for a semi-automatic approach in which research assistants manually examine the documents to identify inconsistencies and harmonize the information. We have established a set of rules to manually transform the original dataset into a consistent format that can be parsed automatically. These rules include separating catalogues with information on multiple main archival persons into separate documents, harmonizing subsection titles, and using "/" to mark line breaks, "#" for general comments, and "##" for comments with additional information. We are also standardizing sections that contain information about correspondence between other actors by introducing a consistent format to replace the diverse formats found throughout the dataset. An edited file excerpt is presented in Figure 5, illustrating an example from Ida Aalberg's catalog. In this example, we have added "/" to mark line breaks, seperated authors with " ; " and added a "#" in front of a comment.

4. Data Model and Harmonization

In order to present the aggregated heterogeneous epistolary datasets in a coherent, unified format, work on developing a harmonizing data model for epistolary metadata collections is undergoing in the project. The CoCo data model builds on international standards like CIDOC CRM[11] [14], Dublin Core[12], and ICA Records in Contexts[13] to promote interoperability. The data model aims to support modeling of the relevant properties of letter metadata that we

[11] https://www.cidoc-crm.org

[12] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[13] https://www.ica.org/en/records-in-contexts-ontology

have collected from the source datasets (see Table 2), to support efficient use of the data in e.g. SPARQL queries and the semantic portal developed during the project.

The current version of the data model includes the most central classes that are Letter, Production, Actor, Place, and Time-Span. Also, provenance (class MetadataRecord) and archival/collection level information (classes Series and Fonds) are included in the data model. During the transformation process, the intermediary RDF format is converted into RDF format corresponding to the CoCo data model. Instances of classes, such as Actor, Place, and Time-Span, are created based on the literal values of the intermediary format.

For representing actors (senders and recipients of letter) in different source datasets, we use an adaptation of the proxy concept from Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[14]. In our case, a Proxy stands for a certain perspective on a person or group in the context of a specific source. In the harmonization process, proxies that are identified (by the future deduplication/disambiguation workflow) to represent the same person or group are connected using a shared instance of the class ProvidedActor. The class is an adaptation of the Europeana Data Model's[15] class ProvidedCHO (Provided Cultural Heritage Object). In Europeana, a ProvidedCHO "represents the Cultural Heritage Object that Europeana collects descriptions about".

The actor data is enriched by linking it to external databases like Wikidata and the Finnish AcademySampo [15] and BiographySampo [16]. 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.

5. Discussion and Conclusion

Recent digital history discussions emphasise the moral or the ethical side central to the use of big data resources. [2, 17] Also, a distinction has been made between technical and ethical data work. [2]. The data transformation pipeline described above, with its different stages of manual, semi-manual and automated processing, can be understood as the technical side of such a project, sometimes referred to as "data cleaning". However, Katie Rawson and Trevor Muñoz have persuasively argued that we should avoid using this phrase. According to them, its (often slightly) offhand use implies that researchers regard the time-consuming data processing as having no tangible impact on the value of the research findings and therefore its detailed description has no relevance. Rawson and Muñoz use phrases such as "critically attuned data work" that enables us to see "the messiness of data not as a block to scalability, but as a vital feature of the world that our data represents and from which it emerges." [18]

The great variety of formatting and the combination of scalable and nonscalable elements in the Word files seemed at first to be a real obstacle for the data processing. However, we gradually realised that without the existence of this diversity we would not understand our data – also the parts we received for example in CSV format – as well as we currently do.

[14] https://www.openarchives.org/ore/

[15] https://pro.europeana.eu/page/edm-documentation

Moreover, it turned out that what we are doing is not merely ”cleaning up” other peoples’ mess as efficiently as possible. As we work through or with the data, we genuinely seek to understand the specificities of the data and consequently try to harmonise it in ways that do not lose the specificity of each correspondence. In this process, we create a new dataset (archive) with its own, regulated vocabulary.

The premises of the ethical data work ran parallel to Rawson’s and Muñoz’s discussions. According to the pivotal *The Network Turn* by Ahnert *et al.*, ethical data work aims at revealing gaps, biases and holes in data sets. The authors suggest that we should look at our data as a perspective, not as a bias. [2] Ethical data work demands a high-standard documentation of all the measures and the whole process of data work done in the project. We must be open towards the future users of our portal and data, and provide them with the information on the metadata that we have and how we have processed them. It also entails scrutinizing the data in the context of participating CH organisations and their collection histories and policies.

One interesting humanistic approach to discuss this (serendipitous and active) selection process is to frame it in terms of cultural heritage. When we do not reduce “cultural heritage” merely to its material manifestations but rather understand it as a dynamic process of making and becoming, we can conceptualize the analogue and digital cataloguing and describing processes as acts of active heritagization. Thus, the production of metadata constitutes an integral part of the discursive framework of values, meanings and relevance that define institutional heritage preservation. 19th-century letters have been potentially heritagizied when they have been included in the collections of CH organizations, and actively heritagizied when they have been provided with metadata. [19] One could perhaps argue that the work described in this paper adds yet another layer to this process. It re-heritagizes those epistolary fonds included in our dataset. They will become available and visible in a way that would not be possible in an analogue archive, and interwoven into the fabric of other – digitized, laboriously yet respectfully processed and connected – cultural heritage.

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, pp. 415–423. 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.

[3] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. Van de Sompel, The europeana data model (edm), in: World Library and Information Congress: 76th IFLA general conference and assembly, volume 10, IFLA, 2010, p. 15.

[4] W. Ravenek, C. van den Heuvel, G. Gerritsen, The epistolarium: origins and techniques, CLARIN in the Low Countries (2017) 317–323. URL: https://doi.org/10.5334/bbi.26.

[5] 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.

[6] 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. URL: http://doi.org/10.3366/ijhac.2015.0140.

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

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

[9] O. Bruneau, N. Lasolle, J. Lieber, E. Nauer, S. Pavlova, L. Rollet, Applying and developing semantic web technologies for exploiting a corpus in history of science: The case study of the henri poincaré correspondence, Semantic Web 12 (2021) 359–378.

[10] M. Silfverberg, T. Ruokolainen, K. Lindén, M. Kurimo, FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish, Language Resources and Evaluation 50 (2016) 863–878.

[11] T. A. Pirinen, Omorfi—free and open source morphological lexical database for finnish, in: Proceedings of the 20th Nordic conference of computational linguistics (NODALIDA 2015), 2015, pp. 313–315.

[12] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen, E. Hyvönen, Automatic annotation service appi: Named entity linking in legal domain, in: A. Harth, V. Presutti, R. Troncy, M. Acosta, A. Polleres, J. D. Fernández, J. Xavier Parreira, O. Hartig, K. Hose, M. Cochez (Eds.), The Semantic Web: ESWC 2020 Satellite Events, volume 12124 of *Lecture Notes in Computer Science*, Springer-Verlag, 2020, pp. 208–213. URL: https://doi.org/10.1007/978-3-030-62327-2_36. doi:10.1007/978-3-030-62327-2_36.

[13] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, CoRR abs/1912.07076 (2019). URL: http://arxiv.org/abs/1912.07076. arXiv:1912.07076.

[14] M. Doerr, The CIDOC CRM—an ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92.

[15] P. Leskinen, H. Rantala, E. Hyvönen, Analyzing the lives of finnish academic people 1640–1899 in nordic and baltic countries: Academysampo data service and portal, in: DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference, CEUR Workshop Proceedings, long papers, Vol. 3232, 2022, pp. 94–108. URL: http://ceur-ws.org/Vol-3232/paper07.pdf.

[16] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case national biography of finland, Semantic

Web – Interoperability, Usability, Applicability 14 (2023) 385–419. URL: https://doi.org/10.3233/SW-222887.

[17] W. Kansteiner, Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent?, History and Theory 61 (2022) 119–133.

[18] K. Rawson, T. Muñoz, Against cleaning, in: M. K. Gold, L. F. Klein (Eds.), Debates in the Digital Humanities 2019, University of Minnesota Press, 2019, pp. 279–292. URL: https://doi.org/10.5749/j.ctvg251hk.26.

[19] J. Enqvist, I. Pikkanen, Kirjeluettelot kulttuuriperintönä ja tutkimusaineistona: metadatan mahdollisuudet digitaalisen käänteen jälkeen (2023/2024). Accepted, forthcoming in 2023/2024.