Integrating Historical Person Registers as Linked Open Data in the WarSampo Knowledge Graph

Mikko Koho\textsuperscript{1,2}, Petri Leskinen\textsuperscript{1}, and Eero Hyvönen\textsuperscript{1,2}

\textsuperscript{1} Semantic Computing Research Group (SeCo), Aalto University, Finland
\textsuperscript{2} HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland

Abstract. Semantic data integration from heterogeneous, distributed data silos enables Digital Humanities research and application development employing a larger, mutually enriched and interlinked knowledge graph. However, data integration is challenging, involving aligning the data models and reconciling the concepts and named entities, such as persons and places. This paper concerns the entity reconciliation of person entities in military historical person registers for semantic data integration. A probabilistic record linkage process is presented to reconcile person references in different person registers with structured metadata into a single knowledge graph. The process was applied to reconcile three person registers of the popular semantic portal “WarSampo – Finnish World War 2 on the Semantic Web”. The registers contain detailed information about some 100 000 people and are individually maintained by domain experts. This sets demands on the integration process to be automated and adaptable to changes in the registers. An evaluation of the record linkage results is promising, and provides some insight into military person register reconciliation in general.

1 Introduction

A way to enhance our understanding about history is to integrate data from complementary information sources in an interoperable way. Semantic heterogeneity is an obstacle to data integration, which can be solved by the process of semantic reconciliation that makes datasets mutually interoperable [4]. Here Named Entity Linking is often useful, i.e., the task of automatically disambiguating and linking the mentions of entity names to entities in knowledge bases. A related challenging problem is Record Linkage (RL) [6,2] where the goal is to find matching structured data records between heterogeneous databases. A typical application scenario is matching person records in different person registers, which contain structured data about some same persons, but are expressed using different metadata schemas and notations. Using RL, richer global descriptions of persons can be created based on local datasets.

This paper concerns the problem of entity reconciliation and RL of person entities in person registers. As a case study, three complementary datasets about some 100 000 Finnish Second World War soldiers in WarSampo [7,10] are considered. A probabilistic record linkage [6] solution for linking and aggregating the person data from multiple datasets is presented, as well as promising evaluation results. The key idea is to assign weights to various comparisons of metadata fields between person registers. The
weights can tell us what information in the person registers is actually important in disambiguating person records in military historical person registers.

After the links of records between registers are generated, information is aggregated into the actor ontology, which contains the identities and enriched metadata of each person. Integrating the person registers into a single knowledge graph (KG) facilitates biographical and prosopographical research [9].

The WarSampo KG is published as open data3, and is part of the international Linked Open Data Cloud. The WarSampo portal4 [7] demonstrates the usefulness of the resulting KG integrated from various sources. WarSampo uses Linked Data and the event-based CIDOC Conceptual Reference Model (CRM)5 together as a basis for harmonizing various datasets about Finland in the Second World War. The portal provides nine different “perspectives” on the data, each of which provides a customized interactive view of the KG: war ‘Events’, ‘Persons’, ‘Army Units’, ‘Places’, ‘Magazine Articles’, ‘Casualties’, ‘Photographs’, ‘War Cemeteries’, and ‘Prisoners of War’. Since its opening in 2015, WarSampo has been used by more than 660 000 end users, corresponding to more than 10% of the population in Finland.

Related Work. The Historical Population Register (HPR) of Norway is pursuing to cover the country’s whole population in 1800–1964 combining information from church records and censuses, employing RL [12]. The Links project6 has similar goals in the Netherlands aiming to reconstruct all nineteenth and early twentieth century families in the Netherlands based on all civil certificates from this period. Antolie et al. [1] present a case study of integrating Canadian World War I data from three sources: one of soldiers, one of casualties, and a census dataset. The linking from casualties to soldier data uses known names, ranks, and regiments as the key attributes. Cunningham [3] presents integrating a World War I veteran military service record with a census database.

2 Data: WarSampo Person Registers

In WarSampo, information about a single person can be found in multiple person registers, each bringing in some new information about the person. The information found from multiple sources can be combined to create a more complete biography of the person. However, it is challenging to reliably say whether two similar looking person records in different registers refer to the same actual person. The problems arise from the heterogeneity of the metadata, different notations used, the ambiguity of metadata annotations, temporal changes in the persons military status, and errors in data.

The military rank and military unit of a soldier are prone to change in time. There can be different spellings of a name, middle names can be missing, and in Finland many originally foreign surnames of soldiers were translated into Finnish at some time. In practise, the same full name can refer to different people, and different names can refer to the same person. There are currently three different person registers in WarSampo:

3https://doi.org/10.5281/zenodo.3431121
4http://sotasampo.fi/en/
5http://cidoc-crm.org
6Cf. the project homepage https://iisg.amsterdam/en/hsn/projects/links and research papers at https://iisg.amsterdam/en/hsn/projects/links/links-publications.
1. Initial Actor Ontology. The ontology containing 5600 people, and also military units, has been created from various data sources which provide varying levels of detail [11]. For most of the people we have ample biographical metadata, e.g. a person’s full name, the dates and places of birth and death, occupation, and dates of promotions during the military career. However, in some cases the level of detail is not sufficient for disambiguation, e.g., only a surname and military rank may be known.

2. Register of Casualties in the Finnish Wars 1939–1945. The register contains 94700 death records (DR) [8], depicting the status of the person at the time of his/her death. The spreadsheet source data contains detailed information about the known Finnish persons who perished in WW2. There are 32 columns of structured information about each person, with each cell having a single literal value.

3. Register of the Prisoners of War in Soviet Union 1939–1945. The register contains 4450 prisoner records (PR) [9], depicting the status of persons at the time when they were captured. It was published in WarSampo on November 2019. The spreadsheet source data contains mostly very detailed information about each known Finnish prisoner of war. The spreadsheet contains 45 columns of information about each person, gathered from, e.g., various archives. Often a single cell contains multiple values corresponding to information in different sources, following a pre-defined cell formatting. Most of the cells contain well-formed literal values, like the municipality of birth, military rank, and date of returning from captivity.

3 Method: Linking Person Records

The WarSampo KG is built from source datasets using a repeatable data transformation pipeline [10]. In this approach, the domain experts maintain the primary data in the original native format, i.e., typically spreadsheets. When a source dataset is updated, the pipeline can be used to easily recreate the whole KG with the updated data.

The pipeline transforms the source spreadsheets of DRs and PRs into RDF, mapping the columns to RDF properties, with possibly multiple values per property. Automatic probabilistic entity linking processes then link the records to the WarSampo domain ontologies of military ranks, units, occupations, people, and places. Also if the domain ontologies are updated, the whole integration process can be redone to account for the changes in the probabilistic entity linking.

The two record linkage scenarios that are needed to tackle for integrating data from all three person registers are:

**RL1.** DRs (94700 person records) linked with the initial actor ontology (5600 persons)

**RL2.** PRs (4450 person records) linked with Actor ontology + DRs (99667 persons)

The first developed solution, applied in both scenarios, was a deterministic / rules-based RL, in which all person pairs were compared with each other, and scored based on a pre-defined handcrafted formula. This was manually evaluated to provide at least satisfactory results, but as the datasets were being maintained and the ontologies evolving, manually maintaining the scoring formula was decided to be not feasible.

The second solution was to use probabilistic RL [6], with a logistic regression-based machine learning implementation employing the Dedupe Python library [5]. Validated
results from the previous solution are used as training data. Persons with only 3 or less metadata fields for RL are completely skipped. The RL solution is open-source\(^7\), and is used in the transformation processes of the DRs\(^8\) and the PRs\(^9\). A run of the probabilistic RL process completes within a few hours in both of the scenarios on an average desktop computer.

The scoring of possible pairs between the PRs and the persons already integrated to WarSampo, i.e., initial actor ontology and DRs, are performed using the comparisons of properties shown in Table 1. The weighted sum of the individual comparisons is used as a confidence that a given pair of records is a match, i.e., that it refers to the same real world person. The comparisons of type string use hyper-parameter optimization to find the best performing string comparison for the values, i.e., Jaro-Winkler. The intersection comparisons compare the one or more URI values of both records to see if there is a matching URI or not. The date comparisons measure the distance of two dates and the dates are based on CIDOC CRM time-spans, which have separate earliest and latest dates. The numerical comparison is measuring the distance of numerical values.

The military ranks of soldiers can change due to promotions or even demotions. As there has been variance also in the use of the abundant private level ranks, a comparison based on the rank level of each military rank was added. This also addresses the rather permanent separation between enlisted ranks and commissioned officers.

**Aggregating Personal Information.** After the links of records between registers are generated, information is aggregated to the actor ontology, which contains the identities and basic metadata of each actual person, with a data model based on CIDOC CRM. New person instances are created in the actor ontology for the records that didn’t match any existing person. The person records are modeled as instances of CIDOC CRM’s document class, which are linked to the person instances in the actor ontology.

---

\(^7\)https://github.com/SemanticComputing/warsa-linkers
\(^8\)https://github.com/SemanticComputing/Casualty-linking
\(^9\)https://github.com/SemanticComputing/WarPrisoners

---

Table 1. Used metadata comparisons of pairs of people between the registers.

| Property                   | Comparison type | Binary/Continuous variable |
|----------------------------|----------------|----------------------------|
| Given names                | string         | continuous                 |
| Family name                | string         | continuous                 |
| Municipality of birth      | intersection   | binary                     |
| Date of birth [earliest]   | date           | continuous                 |
| Date of birth [latest]     | date           | continuous                 |
| Date of death [earliest]   | date           | continuous                 |
| Date of death [latest]     | date           | continuous                 |
| Municipality of death      | intersection   | binary                     |
| Activity end               | date           | binary                     |
| Military rank              | intersection   | binary                     |
| Military rank level        | numerical      | continuous                 |
| Military unit              | intersection   | binary                     |
| Occupation                 | intersection   | binary                     |
4 Results and Evaluation

The record linkage scenario RL1 results in 620 DRs linked to matching people in the 5611 pre-existing person instances, corresponding to 11% of the people in the actor ontology. For the remaining 94,056 DRs, new person instances are created.

The RL2 scenario results in 1397 person records linked to matching people in the 99,667 pre-existing person instances, corresponding to 31% of the PRs. For 3031 PRs, new person instances are created in the actor ontology.

Comparison Weights. The learned comparison weights depict what information is useful for disambiguating the person records. The weights of the comparisons generally vary a little when new runs on updated data are done, but their general magnitude seems to be stable. For the newest WarSampo data transformation, the comparison weights in the RL2 scenario in descending importance order are: family name (2.3), municipality of birth (2.0), given names (1.4), date of birth earliest (1.2), birth date latest (1.2), military rank (1.0), occupation (0.9), military unit (0.8), military rank level (0.8), municipality of death (0.4). The remaining comparisons have weight under 0.1.

Names, municipality of birth and date of birth are intuitively very important personal details defining a persons identity. As the date of birth is split into two comparisons, it’s overall importance can be summed up to 2.4, making it the single most important metadata field. The summed weight of military rank, 1.8, is higher than that of given names. Military unit is also important, nearly as much as a person’s occupation. Occupation of soldiers probably have not been changing during the war, but what is considered a persons occupation might vary depending on the situation and accountant.

Linking Quality. Due to the mostly rich data of each person contained in the person registers, manual evaluation of found links is usually possible, by examining the data in detail. This enables estimating the RL precision. Recall estimation however, would need manual inspection of a very high amount of possible pairs, of which some do not have much information. Also, the DRs are known to have plenty of errors. Hence, it is in many cases difficult to confidently determine the true negative results.

The precision of the record linkage in both scenarios RL1 and RL2 was manually evaluated to be 1.00, based on randomly selecting 150 links from the total of 620 links for RL1, and 200 links from the total of 1397 links for the RL2. The information on the person records and the person instances was compared, and all of the linked records were interpreted to be depicting the same actual persons with high confidence.

5 Discussion

This paper presented the probabilistic record linkage process used in WarSampo to integrate heterogeneous person registers into a reconciled knowledge graph. The weights of different metadata field comparisons, assigned using logistic regression, shed light into what metadata fields are useful in disambiguating person references in the military history context. Military rank and military unit are both important person details when determining the identity of a person depicted in a person record.

The Persons perspective of the WarSampo portal uses the aggregated person instances and information directly from the linked person records to create a unified view
of all the information of each person. The used approach is scalable and can be further used to integrate more person registers into WarSampo. The approach is applicable also to other studies integrating historical person registers.

Acknowledgements

Our work has been funded by the Association for Cherishing the Memory of the Dead of the War, Teri-Säätiö, Open Science and Research Initiative of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

References

1. Antonie, L., Gadgil, H., Grewal, G., Inwood, K.: Historical Data Integration, a Study of WWI Canadian Soldiers. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). pp. 186–193. IEEE (2016)
2. Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Science & Business Media (2012)
3. Cunningham, A.: After “it’s over over there”: Using record linkage to enable the reconstruction of World War I veterans’ demography from soldiers’ experiences to civilian populations. Historical Methods: A Journal of Quantitative and Interdisciplinary History 51, 1–27 (2018)
4. Gal, A., Anaby-Tavor, A., Trombetta, A., Montesi, D.: A framework for modeling and evaluating automatic semantic reconciliation. The VLDB Journal – The International Journal on Very Large Data Bases 14(1), 50–67 (2005)
5. Gregg, F., Eder, D.: Dedupe (2019), https://github.com/dedupeio/dedupe
6. Gu, L., Baxter, R., Vickers, D., Rainsford, C.: Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences (2003), cMIS Technical Report No. 03/83
7. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer (2016)
8. Koho, M., Hyvönen, E., Heino, E., Tuominen, J., Leskinen, P., Mäkelä, E.: Linked Death - Representing, Publishing, and Using Second World War Death Records as Linked Open Data. In: The Semantic Web: ESWC 2017 Satellite Events, pp. 369–383. Springer (2017)
9. Koho, M., Ikkala, E., Hyvönen, E.: Reassembling the Lives of Finnish Prisoners of the Second World War on the Semantic Web. CEUR Workshop Proceedings (2019), in press
10. Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., Hyvönen, E.: WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. Semantic Web – Interoperability, Usability, Applicability (2020), http://semantic-web-journal.net/content/warsampo-knowledge-graph-finland-second-world-war-linked-open-data, accepted
11. Leskinen, P., Koho, M., Heino, E., Tamper, M., Ikkala, E., Tuominen, J., Mäkelä, E., Hyvönen, E.: Modeling and Using an Actor Ontology of Second World War Military Units and Personnel. In: Proceedings of the 16th International Semantic Web Conference (ISWC 2017). pp. 280–296. Springer (October 2017)
12. Thorvaldsen, G., Andersen, T., Sommerseth, H.L.: Record linkage in the historical population register for Norway. In: Population reconstruction, pp. 155–171. Springer (2015)