Integrating Prisoners of War Dataset into the WarSampo Linked Data Infrastructure

Mikko Koho[1], Erkki Heino[1], Esko Ikkala[1], Eero Hyvönen[1,2], Reijo Nikkilä[3], Tiia Moilanen[3], Katri Miettinen[3], and Pertti Suominen[3]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
[3] The National Prisoners of War Project
http://seco.cs.aalto.fi/projects/sotasampo
http://heldig.fi
https://www.arkisto.fi/fi/kansallisarkisto/hankkeet/sotavankiprojekti

Abstract. One of the great promises of Linked Data and the Semantic Web standards is to provide a shared data infrastructure into which more and more data can be imported and aligned, forming a sustainable, ever growing knowledge graph or linked data cloud, Web of Data. This paper studies and evaluates this idea in the context of the WarSampo Linked Data cloud, providing an infrastructure for data related to the Second World War in Finland. As a case study, a new database of prisoners of war with related contents is transformed into linked data and integrated into WarSampo. Lessons learned are discussed in relation to using traditional data publishing approaches.

1 Introduction

Data about the Second World War (WW2) is heterogeneous and distributed in different organizations and countries. WarSampo [7] provides a novel infrastructure for publishing WW2 data as Linked Open Data. The infrastructure is built to support integrating new datasets into WarSampo, by extending the WarSampo data model, as well as the data content. The idea is to be able to easily extend the database with new digitized war datasets, and link them to existing datasets already in the database. The WarSampo web portal builds upon these interlinked datasets, by providing different perspectives to the whole database as customized web applications. New perspectives can be added easily to provide views to new datasets, or to answer new research questions with existing data.

In the *Linked Data* architecture [5,6], data is presented as graphs, with identifiers for all resources. These identifiers (URIs or IRIs) are used to link pieces of data together in the graph, and also to link to external datasets. The identifiers can be resolved using normal web technologies to fetch the descriptions of resources. Linked Data uses the *RDF* data model[4], in which everything is expressed as triple statements consisting of a subject, predicate, and object. Linked Data that is published openly online, is often referred to as *Linked Open Data*.

[4] https://www.w3.org/RDF/

This paper presents preliminary results of integrating data about the Finnish prisoners of war into the WarSampo infrastructure. This data publishing method is compared to two competing traditional, approaches that could also have been used for the data publication. We evaluate the following three data publishing approaches:

1. **Data files.** Publishing the data as separate data files, e.g., as spreadsheets.
2. **Database & web interface.** Inserting the data into a (relational) database, and creating a web user interface for querying and displaying the data.
3. **Linked Data.** Publishing the data as Linked Data, using the WarSampo infrastructure.

These approaches are compared to each other to find the benefits and drawbacks of each. The dataset is being published using the WarSampo approach, but exists also as spreadsheets, which effectively allows for comparison of these approaches. The Database and web interface option is evaluated based on our experience with these technologies, and also on examination of the online publication of the Finnish WW2 casualties of war[5] database by the National Archives of Finland, a dataset that has also previously been integrated into WarSampo [9].

A new application perspective was created into the WarSampo portal for studying the prisoners of war dataset individually, while also integrating the prisoners into the WarSampo person perspective, and extending person pages to show the new prisoner data. The new perspective facilitates prosopographical study of the prisoners using the whole dataset, or a subset of the data based on user interest.

The core of the new data is a register of Finnish prisoners of war, containing data about some 4 460 prisoners in the Soviet Union as a spreadsheet. Additional information about related prisoner camps and hospitals is contained in separate spreadsheets. The data includes also documents about the prisoners of war to provide additional information.

There are many studies about the Finnish WW2 prisoners of war [11,3,1], but still there is much that is not known. The exact number of prisoners of war is still unknown, as is many details of them. Of some individuals, there is not much else known than their name.

The prisoner data has originally been published in a book [1]. For this study the data has been manually extended, cleaned, and validated extensively. This work was done by domain experts who have inspected large amounts of wartime archives from Finland and the Soviet Union, among other data sources.

The data in the prisoner register has been gathered from several Finnish and Russian archives [1]. Data in different sources can be contradictory, and in order to preserve the provenance information of the pieces of data, the prisoner register contains the data source for each value. Each column can have multiple values, and possibly source information for each value. Multiple values in a spreadsheet cell of the register are transformed into multiple values of an RDF property in the linked data publication.

[5] `http://kronos.narc.fi/menehtyneet/`

The alternative approaches to publishing the prisoner data are discussed in the next two sections, followed by the description and evaluation of publishing the dataset as part of the WarSampo data service (SPARQL endpoint) in the Linked Data Finland service[6] and the semantic portal[7] for end users.

2 Prisoners of War as Data Files

The simplest way of publishing the dataset would be to make it available as spreadsheets in, e.g., CSV format. This enables easy open access to the dataset, but does not facilitate the creation of interactive applications using the data. The data would also not benefit from information in other datasets, as no links to other datasets would exist. The benefits (+) and challenges (-) of this approach are:

+ **Low effort.** Not much work is required to publish the data online.
- **Static data.** The dataset can only be accessed as static files, and not through a programmable interface.
- **No linking.** The data forms a new silo that can not communicate with other datasets.
- **No interactive applications.** The data is in a format that is not feasible to use for creating interactive web applications.

3 Prisoners of War as a Relational Database

In order to manage the data in a structured way, and to develop interactive applications using it, the dataset could be stored in a relational database. This would also enable the use of pre-existing tools and frameworks for said tasks. An *application programming interface (API)* could be developed for programmatic online access to the data, and a web-based application for end-users to browse the data.

However, this approach would still lack a standard way of linking the data to other data sources, and access to the data would be limited by the custom API. The pros and cons are thus:

+ **Existing tools.** Tools and frameworks for creating and managing relational databases and applications built using them are ubiquitous.
- **No linking.** The data still forms a new silo that can not communicate with other datasets.
- **Non-standard data publication.** The dataset can be made available through a custom API, but not freely queryable in a standard way.

[6] `http://ldf.fi`
[7] `http://sotasampo.fi/en`

4 Prisoners of War as Linked Data

Publishing data as Linked Data makes it possible to link individual pieces of information to each other via different relations. Related datasets can be linked to each other if they share same concepts or entities, like persons, military units or places. WarSampo is based on Linked Data, and is published openly via a SPARQL endpoint.

In contrast to the data file or relational database approaches, the linked data approach requires tighter co-operation with the domain experts and data publishers, especially in the creation phase in the life cycle of historical information [2]. The domain experts working in the National Prisoners of War Project looked through various sources and fed the data into spreadsheets. For integrating the prisoner data into the WarSampo infrastructure, the following decisions concerning the structure and contents of the spreadsheets were carried out in advance with the domain experts:

- Choose the values that will be automatically linked to the WarSampo domain ontologies, and separate them into distinct columns.
- Introduce identifiers for entities that appear on multiple spreadsheets.
- Develop a common practice for attaching a source for each value in a spreadsheet cell.
- Decide how to express partially or completely missing information.

After initial versions of the spreadsheets were finished, a pipeline for converting the prisoner data into RDF format and linking them to the WarSampo domain ontologies was created. Source codes for data conversion and linking are available online[8].

Due to lack of user friendly tools for editing the data in RDF format, it was decided that the domain experts kept on working with the spreadsheets that they were accustomed to use. Because of this the pipeline had to be adjusted every time the structure of a spreadsheet was modified. However, running the pipeline regularly fostered iterative development as the domain experts were able to preview how the spreadsheet data is visualized in the WarSampo web portal, and to suggest corrections.

4.1 Linked Data Model

The prisoners of war as Linked Data, created with the data transformation pipeline, is published in the WarSampo SPARQL endpoint[9], in a separate named graph `(http://ldf.fi/warsa/prisoners)`.

We use a simple primary data model for the data in RDF, resembling the original format, in which the data was presented as spreadsheets. The data model is similar to that of the WarSampo casualties [9], and their properties and classes

[8] `https://github.com/SemanticComputing/WarPrisoners`
[9] `http://ldf.fi/warsa/sparql`

have been harmonized using the dumb-down principle of Dublin Core[10], i.e., by using shared super-properties and super-classes where applicable.

WarSampo uses the *CIDOC Conceptual Reference Model (CRM)*[11] as a harmonizing data model. Individual prisoner records are modeled as instances of the CRM document class (`E31_Document`). Each record corresponds to a simple RDF transformation of a single database row, where the properties used correspond to the columns of the table. Using these properties, the data can be shown to the end user in an intuitive way similar to the orginal table.

In addition, this data is then used to create CIDOC CRM descriptions of the actual persons and events, when appropriate. WarSampo person instances [10] are enriched using the prisoner of war documents. New person instances are created for persons that do not already exist in the knowledge base, which is the case for most of the war prisoners.

The Linked Data publication stores source information when present in the original data. There are many ways of presenting this kind of provenance information in RDF [4,12]. The approach used with the prisoners of war dataset is storing source information using RDF reification with the DCMI Metadata Terms[12] property *source*.

Some parts of the data had to be left out of the online data publication due to privacy legislation. This is done automatically based on the dates when prisoners have died, since the legislation does not concern people who have passed away more than 50 years ago. Basic information, like name and date of birth, can be published of all prisoners of war.

4.2 Enriching the Data Using Other WarSampo Datasets

One of the most challenging aspects of the data transformation pipeline is matching the persons from different datasets. The data models and content can be highly varying, and different parts of person data can be missing in some datasets. We were able to link 1 417 prisoners (32% of all prisoners) to persons that already existed in the WarSampo actor ontology [10]. Most of these people originate from the WarSampo casualties dataset.

The prisoners' rank and military unit information were linked to the corresponding WarSampo datasets, of which 99% and 90% were linked successfully, respectively. Successful linking allows for enriching all of these datasets with each other's information content.

4.3 Prisoners of War as Part of WarSampo

The new prisoner perspective of WarSampo uses SPARQL Faceter [8] to provide a faceted search interface for the prisoner data. The prisoner perspective

[10] `http://dublincore.org/usage/documents/principles/`
[11] `http://cidoc-crm.org`
[12] `http://dublincore.org/documents/dcmi-terms/`

application is open-source, and available online[13]. The prisoner perspective in WarSampo will be opened to the public in March 2018. The interface facilitates prosopographical study of the prisoners by enabling the grouping of people based on various facets, and visualizing the results in different ways. Further details about each individual prisoner are made available through links to the WarSampo person perspective.

The WarSampo person perspective shows a biographical view of a single person by bringing together all information that exists about the person — including the information in the prisoner dataset. Each person's page shows the basic information about a person, while also showing the prisoner record, and additional documents, like interrogation sheets, pictures, and videos that exist about that individual.

A major benefit of the Linked Data approach is that the SPARQL endpoint allows to find answers to many complex research questions by querying the data directly. Asking complex queries, however, requires a person to be familiar with the SPARQL syntax. We can, for example, easily find out what are the battles of the Winter War, where the most soldiers have become prisoners, by exploiting the unit links of the prisoner records, and the existing data links in WarSampo. A single SPARQL query[14] can reveal this information. The result of this query shows us what military unit was involved in a battle, how many prisoners of war were taken during a battle, how many prisoners of war were taken from the unit totally, and how many soldiers perished during the battle.

The most prisoners were taken during the battle of *Leipäsuo*. This also accounts for 86% of the prisoners of war of the involved military unit (*JR9*). This information also enriches the prisoners of war data, since prisoners taken from *JR9* were not previously annotated with any place of capture — now due to linking we know where they were at the time. The five Winter War battles found in WarSampo in which the most prisoners were taken are the following:

1. Leipäsuo, military unit *JR 9* (4.1.1940 - 14.2.1940, 25 captured, 215 dead).
2. Summa, military unit *JR 13* (11.2.1940 - 16.2.1940, 22 captured, 74 dead).
3. Summa defensive battle, military unit *JR 14* (12.2.1940 - 15.2.1940, 15 captured, 52 dead).
4. Defensive battle at Laatokka-Syskyjärvi, military unit *JR 37* (11.12.1939 - 28.12.1939, 14 captured, 37 dead).
5. Summa defensive battle, military unit *JR 14* (14.2.1940 - 16.2.1940, 13 captured, 27 dead).

4.4 Evaluation

Benefits and challenges of the **Linked Data** approach are:

+ **Availability.** The data is published on an open standardized API for others to use. Resolvable identifiers allow anyone to look up entities using normal web technologies.

[13] `https://github.com/SemanticComputing/prisoners-demo`
[14] The query is available at `http://yasgui.org/short/Hy2bTFOCb`

+ **Identifiers.** Identifiers are used for all information; this enables the linking of the pieces of data. Also future datasets can refer to the same identifiers.
+ **Interlinking.** The data can be enriched from multiple related datasets.
+ **Web user-interfaces.** It is easy to implement user-friendly web perspectives that dynamically use the data via a SPARQL API.
+ **Validation.** The data transformation process validates the structure of the data, and can be used to find structural errors easily.
- **More work needed.** Creating a new RDF data model, transforming the data into RDF, interlinking to other datasets, and extending the web portal requires more work.
- **Coordination.** Data linking requires more coordination between data producers and data publishers.
- **Slower querying.** Querying data in the RDF data model usually results in higher response times than data in simpler formats.

5 Conclusion

This paper overviewed a case study of publishing Finnish prisoners of war content as linked open data, using the WarSampo infrastructure. We discussed the pros and cons of this approach in relation to two competing traditional approaches.

A significant difference between the traditional data publishing approaches (data file, relational database) and the Linked Data approach is that the traditional approaches allow for more straightforward workflows. Data cleaning and validation is not as critical, and the data gathered by domain experts can be converted into a desired format using simple scripts that do not necessarily modify the original data in any way. The Linked Data approach, on the other hand, always requires careful inspection of the data provided by the domain experts.

This paper illustrated that linking the data to the domain ontologies is one key distinction between the traditional and Linked data approaches. The linking process includes demanding tasks, such as choosing which part of the data is suitable for linking, cleaning and harmonizing the data, and finally developing automatic or manual workflows for data linking.

Our experiment suggests that the data linking process should take place as early as possible in the lifecycle of publishing historical information, preferably in the data creation phase, in order to maximize precision and recall. With this project it was possible to facilitate the linking process with the domain experts during the data creation phase, but often this is not the case. Applying Linked Data publishing principles to a dataset that has been previously published as a data file or database requires either error-prone interpretation of the original data, or co-operation with the domain experts that created the original data.

Acknowledgements

Our work was funded by the Association for Cherishing the Memory of the Dead of the War[15], Open Science and Research Initiative[16] of the Finnish Ministry of Education and Culture, the Finnish Cultural Foundation, and the Academy of Finland.

References

1. Alava, T., Frolov, D., Nikkilä, R.: Rukiver. Suomalaiset sotavangit Neuvostoliitossa. Helsinki: Edita (2003)
2. Boonstra, O., Breure, L., Doorn, P.: Past, present and future of historical information science. Historical Social Research 29(2), 4–132 (2004)
3. Frolov, D.: Sotavankina Neuvostoliitossa: suomalaiset NKVD: n leireissä talvi- ja jatkosodan aikana. Suomalaisen Kirjallisuuden Seura (2004)
4. Hartig, O.: Provenance information in the web of data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web. No. 538 in CEUR Workshop Proceedings (2009), http://ceur-ws.org/Vol-538/ldow2009_paper18.pdf
5. Heath, T., Bizer, C.: Linked Data: Evolving the web into a global data space. Synthesis Lectures on The Semantic Web: Theory and Technology, Morgan & Claypool Publishers, Palo Alto, USA (2011)
6. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers, Palo Alto, USA (2012)
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-Verlag (2016)
8. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. No. 1615, CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
9. 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: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds.) The Semantic Web: ESWC 2017 Satellite Events. pp. 369–383. Springer-Verlag (2017), https://doi.org/10.1007/978-3-319-70407-4_45
10. 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). Springer-Verlag (October 2017)
11. Malmi, T.: Jatkosodan suomalaiset sotavangit neuvostojärjestelmässä. Licenciate thesis, University of Tampere (1996)
12. Zhao, J., Bizer, C., Gil, A., Missier, P., Sahoo, S.: Provenance requirements for the next version of RDF. In: Proceedings of the W3C Workshop – RDF Next Steps. W3C, https://www.w3.org/2009/12/rdf-ws/papers/ws08

[15] http://www.sotavainajat.net/in_english
[16] http://openscience.fi/