How to Maintain a Linked Data Cloud in a Deployed Semantic Portal

Mikko Koho[1], Esko Ikkala[1], Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland and
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
`http://seco.cs.aalto.fi, http://heldig.fi`

Problem Statement Plenty of data exists about the Second World War (WW2), but the data is scattered in distributed databases, written in multiple languages, and recorded in heterogeneous formats. Scenarios of this kind are difficult for cultural heritage organizations and companies, where the data is fragmented and all the different pieces are related to each other. A common infrastructure is needed for the data, along with interfaces that allow citizens to browse and study the data, and Digital Humanities researchers to use the data easily for research. WW2 information is of great interest not only to historians, but to potentially hundreds of millions of citizens globally whose relatives participated in the war, creating a global shared trauma.

Data values in different datasets are not directly interoperable as the metadata are not harmonized using shared vocabularies. Using shared vocabularies, however, creates new kinds of challenges as changes in the shared vocabularies need to be handled in the datasets using them [2]. For the shared vocabularies to be sustainable, they need to be maintained by a responsible administrative authority, in collaboration with the actual users of the vocabularies. The vocabularies need to be expanded dynamically when new values are needed in them.

This kind of a complex scenario makes a promising use case for Linked Data (LD), as various pieces of data can be connected in a flexible way. However, maintaining LD brings in new technical challenges, like data transformation, entity linking, and change propagation between graphs. The LD field is not mature enough to provide tools that could be easily used by the domain experts for the maintenance of Linked Data Clouds (LDC), dynamic systems of multiple large interlinked graphs.

The WarSampo LDC Published in 2015, WarSampo [1] provides a dynamic ontology infrastructure for serving WW2 data as Linked Open Data (LOD), and a growing collection of datasets, totaling ca 12 million triples. The infrastructure is built to support integrating new datasets into WarSampo, by extending both the domain ontologies and the data graphs. The work is done as a collaboration between LD researchers and organizations possessing WW2 related data, such as the National Archives of Finland, the National Land Survey of Finland, and the Finnish Defense Forces. WarSampo is a part of the global international LOD cloud and was awarded the LODLAM Challenge Open Data Prize in 2017. The WarSampo semantic portal[3] builds upon the WarSampo data,

[3] `https://www.sotasampo.fi/en/`

which is served on an open SPARQL endpoint[4]. The portal provides different perspectives to the knowledge base as customized web applications.

WarSampo was recently extended by a dataset of hundreds of war cemeteries and thousands of photographs of them, and then by another dataset of about 4450 Finnish prisoners of war. The War Cemetery perspective was published in November 2017 and got 57.000 users in one week, due to media coverage, which set new demands for scalability. The Prisoners of War perspective is going to be published in late 2018. In total, the WarSampo semantic portal was used by 130 000 users in 2017. The data service is based on a Fuseki SPARQL server on a scalable cloud computing platform. The semantic portal is capable of handling at least hundreds of concurrent users, accessing the data directly via SPARQL.

Lessons Learned Due to the interlinked graphs, the maintenance of the data in RDF format is difficult. For example, the person instances are linked, directly or indirectly, to everything in WarSampo. Modeling even just a person's basic information entails e.g. multiple events, such as birth, death, promotions, and so on. So instead, the domain experts maintain the datasets in their native formats (e.g. spreadsheets), which can then be reintegrated into WarSampo, as needed. Reintegration is also used for solving the change propagations from one graph to another. Sustainable long-term maintenance should be facilitated by an administrative cultural heritage organization with domain knowledge, which would need more mature and easy-to-use applications for data transformation, entity linking, and LD maintenance.

With the LDC approach, we can harmonize inconsistent values in the data integration phase, and mint URIs to new entities in DOs as needed. The LDC can be analyzed and visualized as a whole, instead of focusing on individual datasets. E.g., we can visualize the whole LDC on a map as a function of time. There are other online services based on Finnish WW2 data, without semantic technologies, such as the Casualties of War service of the National Archives and Sotapolku[5]. The services suffer from the lack of harmonization of data values, missing linkage between entities, and lack of public data APIs.

Although integrating data into the LDC is more laborious than simpler ways of publishing the data online, the result is an interlinked knowledge base, where the graphs enrich each other through the linking, creating a whole that is greater than the sum of its parts.

References

1. 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)
2. Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D., Sabou, M.: Ontology evolution: a process-centric survey. The Knowledge Engineering Review 30(1), 45–75 (2015)

[4] `http://ldf.fi/warsa/sparql`

[5] `http://kronos.narc.fi/menehtyneet/`, `https://www.sotapolku.fi`