Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Knowledge Discovery

Eero Hyvönen *,
[a] *Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, and Department of Computer Science, Aalto University, Finland*
E-mail: eero.hyvonen@aalto.fi

Abstract. This paper envisions and discusses a shift of focus in research on Cultural Heritage semantic portals, based on Linked Data. While ten years ago the research focus in semantic portal development was on data harmonization, aggregation, search, and browsing ("1st generation systems"), the rise of Digital Humanities research is shifting the focus on providing the user with integrated tools for solving research problems in interactive ways ("2nd generation systems"). This trend sets new challenges for both computer scientists and humanist researchers.

Keywords: Digital Humanities, Semantic portals, Data analysis

1. Introduction

Cultural Heritage (CH) has become a most active area of application of Linked Data and Semantic Web (SW) technologies [1]. Large amounts of CH content and metadata about it are available openly for research and public use based on collections in museums, libraries, archives, and media organizations. Data has been aggregated in large national and international repositories, web services, and portals such as Europeana[1] and Digital Public Library of America[2], and forms a substantial part of DBpedia[3], and Wikidata[4]. From a SW research point of view, the CH data provides interesting challenges: the data is syntactically heterogeneous (text, images, sound, videos, and structured data in different forms), semantically rich covering all aspects of life in different times and places, interlinked across different data sources, multilingual, often incomplete, imprecise, uncertain, or fuzzy due to the nature of history, and distributed in different countries and databases.

1.1. First Generation Semantic CH Portals

As a result, SW research in CH has been initially focused on issues related to syntactic and semantic interoperability and data aggregation. A great deal of work has been devoted in developing metadata standards and data models for harmonizing data, including application agnostic W3C standards (RDF, OWL, SKOS, etc.), document centric models, such as Dublin Core and its dumb down principle, and event centric models for data harmonization on a more fundamental level, such as CIDOC CRM[5] and its extensions for museums, and IFLA Library Reference Model (LRM)[6] in libraries. Based on harmonized data, published in a SPARQL endpoint or via other APIs, a variety of semantic portals have been created. The main use case has been providing the user with enhanced information

*Corresponding author. E-mail: eero.hyvonen@aalto.fi.
[1]http://europeana.eu
[2]https://dp.la/
[3]http://dbpedia.org
[4]http://wikidata.org
[5]http://cidoc-crm.org
[6]https://www.ifla.org/publications/node/11412

retrieval (IR) facilities [2], such as faceted search [3], semantic search, entity search, and semantic recommendation systems [4] for exploring the data in intelligent ways. In the following, CH search systems based on harmonized aggregated linked data will be *1st generation CH semantic portals*.

1.2. Second Generation Semantic CH Portals

As more and more harmonized aggregated linked datasets are available, the time has come to take a next step forward to *2nd generation CH semantic portals*. The novelty of such systems is to provide the user with tools for solving Digital Humanities (DH) [5] research problems, not only tools for searching and browsing the data. In DH, a key goal is to use computational methods for humanist and social science problems using large datasets that have become available. In this context Big Data means data that is too big and complex to be analyzed by close reading [6]. A variety of technologies have been developed for such tasks, such as sentiment analysis [7], topic modeling [8], and network analysis [9], in addition to traditional and novel statistical analysis methods, such as deep learning. There are lots of tools available for these tasks, but a major problem here is that using them typically requires technical expertise and skills not common among the humanist researchers. Furthermore, the tools are not integrated with Linked Data formats and data services, and there is the burden of transforming and transporting linked data into forms required by the different data analysis tools.

This paper argues that semantic CH portals, focusing usually on semantic search and browsing, should extend their services also to data analytic services and tools needed in DH research and applications. To support the argument, I reflect on my own experiences in 10 years in developing the "Sampo" series of semantic portals [10], starting from the portal CultureSampo (2008) [11] and ending up with BiographySampo (2018) [12] that has a particular focus on services for Digital Humanities.

2. Sampo Model for Publishing CH Linked Data

The ideas of the Semantic Web and Linked Data can be applied to address the problems of semantic data interoperability and distributed content creation at the same time, as depicted in Fig. 1. Here the publication system is illustrated by a circle. A shared semantic ontology infrastructure is situated in the middle. It includes mutually aligned metadata and shared domain ontologies, modeled using SW standards. If content providers outside of the circle provide the system with metadata about CH, the data is automatically linked and enriched with each other and forms a knowledge graph.

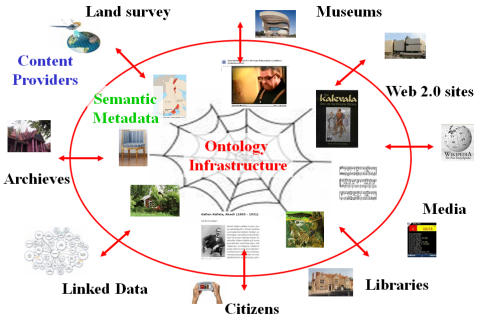

Fig. 1. Sampo model for Linked Data publishing is based on a shared ontology infrastructure in the middle.

For example, if metadata about a painting created by Picasso comes from an art museum, it can be enriched (linked) with, e.g., biographies from Wikipedia and other sources, photos taken of Picasso, information about his wives, books in a library describing his works of art, related exhibitions open in museums, and so on. At the same time, the contents of any organization in the portal having Picasso related material get enriched by the metadata of the new artwork entered in the system. This is a win-win business model for everybody to join such a system; collaboration pays off.

Combining the infrastructure with the idea of decoupling the data services for machines from the applications for the human user creates a model for building collaborative Semantic Web applications. I call this whole the Sampo model. The model has been developed and tested in a series of several practical case studies [10].

3. Tooling for Digital Humanities

Problem solving in DH often has two phases, as in the prosopographical research method [13, p. 47]: First, a target group of entities in the data is selected that share desired characteristics for solving the research question at hand (in the case of prosopography, a people group is selected). Second, the target group is analyzed, and possibly compared with other groups, in order to solve the research question.

The analysis in DH is typically done partly by the machine, partly by the human. In visualizations, such as maps, timelines, and networks, the machine presents target data in a form from which the human user is able to make interpretations more easily. In statistic charts, such as pie charts, line charts, and histograms are used. Another type of tooling is network analysis [14], where different kind of connections between entities, such as family relations between persons or references between texts can be represented as graphs for visual inspection and mathematical analysis. In data-analysis and knowledge discovery statistical or other patterns of data are searched for in order to find "interesting", serendipitous [15] new knowledge. Techniques such as topic modelling [8] fall in this category. The results also here typically need human interpretation, as statistical methods are usually unable to explain their results. In knowledge-based systems, knowledge structures can be used for this.

Many of the methods and tools above are well-defined and domain independent, and there are lots software packages available for using them, such as Gephi[7], R [16], and various Python libraries. However, each of them have their own input formats and user interfaces, and need specific skills from the user. Furthermore, visualizations are crafted case by case; tools for formulating, adjusting, and comparing them in some general ways would be helpful for the user.

4. Case Study: From CultureSampo to BiographySampo

In the case of the first Sampo systems CultureSampo (2008), TravelSampo (2011), BookSampo (2011), and WarSampo (2015) a key element in the user interfaces is faceted search, where search results can be filtered out in flexible ways, and then browsed. They are examples of 1st generation systems. However, first experiments on providing the end-user with DH tooling were done in WarSampo regarding the fairly homogeneus and complete datasets of death records and war cemeteries. A more systematic approach was taken in BiographySampo (2018), where a complete tool set for prosopographical research was designed and integrated in the system based on a SPARQL endpoint. In below, this system is explained using examples in order to illustrate the idea of the 2nd generation CH portal.

[7]https://gephi.org/

Biography studies life stories of particular people of significance, with the aim of getting a better understanding of their personality and actions, e.g., to understand their motives. [17] Biography research on top of a dataset of biographies can be supported by a 1st generation CH system, where data of interest can be searched for using a search engine. In BiographySampo, faceted search is used for this.

In contrast to biography, the focus of *prosopography* is to study life histories of groups of people in order to find out some kind of commonness or average in them. [13] For example, the research question may be to find out what happened to the students of a school before the World War II in terms of social ranking, employment, or military involvement after their graduation. To support prosopography, a 2nd generation CH application is needed. Filtering out the target group is not enough but tools and visualizations are needed for analyzing it, too. In BiographySampo they can be applied not only to one target group but also to two parallel groups in order to compare them. For example, Fig. 2 compares the life charts of Finnish generals and admirals in the Russian armed forces in 1809–1917 when Finland was an autonomous Grand Duchy within the Russian Empire (on the left) with the members of the Finnish clergy (1800–1920) (on the right). With a few selections from the facets the user can see that, for some reason, quite a few officers moved to Southern Europe when they retired (like retirees today) while the Lutheran ministers tended to stay in Finland.

In the same way, the statistical application perspective includes histograms showing various numeric value distributions of the members of the group, e.g., their ages, number of spouses and children, and pie charts visualizing proportional distributions of professions, societal domains, and working organizations. The networks perspective is based on the same idea of visualizing and studying networks among target groups filtered out using facets. The networks are based on the reference links between the biographies, either handmade or based on automatically detected mentions. The depth of the networks can be controlled by limiting the number of links, and coloring of the nodes can be based on the gender or societal domain of the person (e.g., military, medical, business, music, etc.).

To utilize reasoning and knowledge discovery, yet another application perspective for finding "interesting/serendi-pitous" [15] connections in the biographical knowledge graph was created. This application idea is related to relational search [18, 19]. However, in our

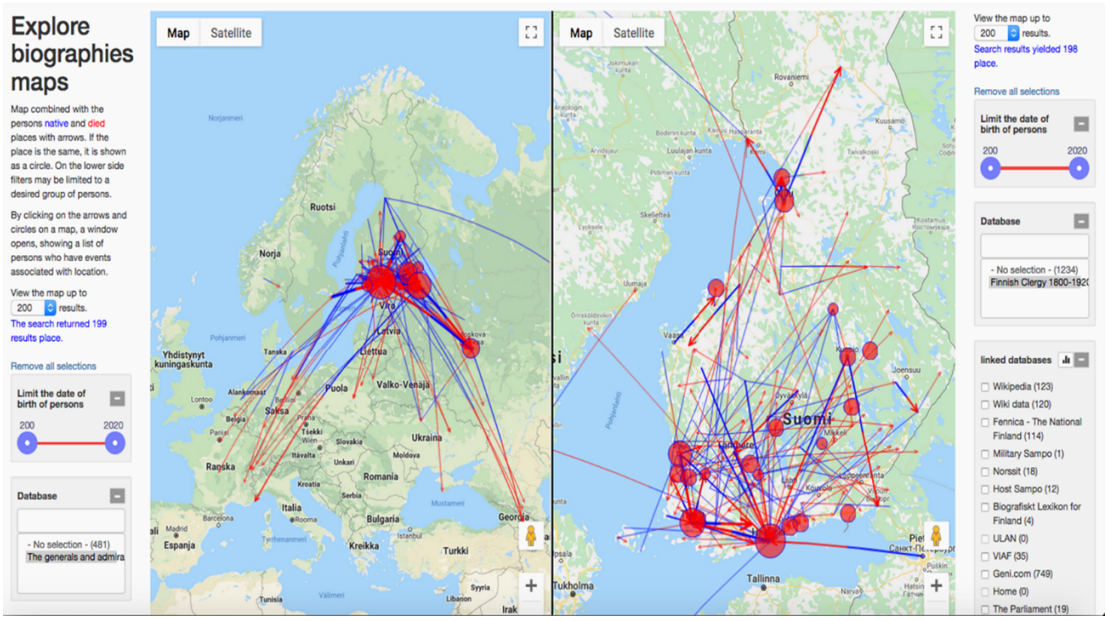

Fig. 2. Comparing the life charts of two target groups, admirals and generals (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).

case a new knowledge-based approach was developed to find out in what ways (groups of) people are related to places and areas. This method, described in more detail in [20], rules out non-sense relations effectively and is able to create natural language explanations for the connections. The queries are formulated and the problems are solved using faceted search. For example, the query ”How are Finnish artists related to Italy?” is solved by selecting ”Italy” from the place facet and ”artist” from the profession facet. The results include connections of different types (that could be filtered in another facet), e.g., that ”Elin Danielson-Gambogi received in 1899 the Florence City Art Award” and ”Robert Ekman created in 1844 the painting ’Landscape in Subiaco’ depicting a place in Italy”.

Finally, the biographies can be analyzed by using linguistic analysis. For example, it turns out that the biographies of female Members of the Parliament (MP) frequently contain words ”family” and ”child”, but these words are seldom used in the biographies of male MPs. The analyses are based on the linguistic knowledge graph of the texts.

5. Conclusions

This paper envisioned and discussed an emerging major trend in semantic portals for CH: while the 1st generation systems provided the end user search and browsing facilities on top of a linked data service (SPARQL endpoint), the 2nd generation systems provide the user also with tools for solving DH problems. Examples of this include, in addition to BiographySampo, e.g., Six Degrees of Francis Bacon [21] and Epistolarium[8] and many others. A typical use pattern here is to first filter out target data of interest and then apply various data analysis and knowledge discovery tools on the target. In the presented case example BiographySampo, faceted search is used for filtering out data and various seamlessly integrated tools are ready to be used interactively by the DH researcher on top of a SPARQL endpoint.

The shift of focus from data publishing to data analysis and tooling for DH brings in novel research challenges in, e.g., knowledge extraction, data visualization, machine learning, and knowledge discovery. In addition, interpreting the results typically requires a great deal of domain knowledge and understanding the underlying algorithms and characteristics of the data,

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

such as its modeling principles, completeness, uncertainty, and fuzziness. Using advanced computational tools in humanities and social sciences raises the demand for source criticism on a new, higher level.

References

[1] E. Hyvönen, *Publishing and using cultural heritage linked data on the semantic web*, Morgan & Claypool, Palo Alto, CA, 2012.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval (2nd Edition)*, Addison-Wesley, New York, 2011.

[3] D. Tunkelang, Faceted search, *Synthesis Lectures on Information Concepts, Retrieval, and Services* **1**(1) (2009), 1–80.

[4] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, *Recommender Systems. An introduction*, Cambridge University Press, Cambridge, UK, 2011.

[5] E. Gardiner and R.G. Musto, *The Digital Humanities: A Primer for Students and Scholars*, Cambridge University Press, New York, NY, USA, 2015.

[6] K. Shultz, What Is Distant Reading?, *New York Times* (June, 24, 2011).

[7] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, Palo Alto, CA, 2012.

[8] M.R. Brett, Topic Modeling: A Basic Introduction, *Journal of Digital Humanities* **2**(1) (2012). http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.

[9] E. Otte and R. Rousseau, Social network analysis: a powerful strategy, also for the information sciences, *Journal of information Science* **28**(6) (2002), 441–453.

[10] E. Hyvönen, Cultural Heritage Linked Data on the Semantic Web: Three Case Studies Using the Sampo Model, in: *VIII Encounter of Documentation Centres of Contemporary Art: Open Linked Data and Integral Management of Information in Cultural Centres Artium, Vitoria-Gasteiz, Spain, October 19-20, 2016*, 2016.

[11] E. Mäkelä, T. Ruotsalo and Hyvönen, How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo, *Semantic Web Journal* **3**(1) (2012), 85–109.

[12] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen and K. Keravuori, BiographySampo - Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, in: *Proceedings of the 16th Extendwed Semantic Web Conference (ESWC 2019)*, Springer-Verlag, 2019, Accepted.

[13] K. Verboven, M. Carlier and J. Dumolyn, A short manual to the art of prosopography, in: *Prosopography approaches and applications. A handbook*, Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70.

[14] M. Newman, *Networks*, Oxford university press, 2018.

[15] R.S. Aylett, D.S. Bental, R. Stewart, J. Forth and G.Wiggins, Supporting Serendipitous Discovery, in: *Digital Futures (Third Annual Digital Economy Conference), 23–25 October, 2012, Aberdeen, UK*, 2012.

[16] A. Field, J. Miles and Z. Field, *Discovering Statistics Using R*, SAGE Publications Inc., USA, 2015.

[17] B. Roberts, *Biographical Research*, Understanding social research, Open University Press, 2002. ISBN 9780335202867.

[18] S. Lohmann, P. Heim, T. Stegemann and J. Ziegler, The RelFinder User Interface: Interactive Exploration of Relationships between Objects of Interest, in: *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI 2010)*, ACM, 2010, pp. 421–422.

[19] G. Tartari and A. Hogan, WiSP: Weighted Shortest Paths for RDF Graphs, in: *Proceedings of VOILA 2018*, CEUR Workshop Proc., Vol. 2187, 2018, pp. 37–52.

[20] E. Hyvönen and H. Rantala, Knowledge-based Relation Discovery in Cultural Heritage Knowledge Graphs, in: *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference (DHN 2019).*, CEUR Workshop Proceedings, 2019.

[21] C.N. Warren, D. Shore, J. Otis, L. Wang, M. Finegold and C. Shalizi, Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks., *DHQ: Digital Humanities Quarterly* **10**(3) (2016).