Digital Humanities on the Semantic Web: Sampo Model and Portal Series

Eero Hyvönen
Aalto University, Department of Computer Science, Finland and
University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland
Semantic Computing Research Group (SeCo) (http://seco.cs.aalto.fi)
E-mail: eero.hyvonen@aalto.fi

Abstract.
Cultural heritage (CH) contents are typically strongly interlinked, but published in heterogeneous, distributed local data silos, making it difficult to utilize the data on a global level. Furthermore, the content is usually available only for humans to read, and not as data for Digital Humanities (DH) analyses and application development. This application report paper addresses these problems by presenting a collaborative publication model for CH Linked Data and six design principles for creating shared data services and semantic portals for DH research and applications. This *Sampo model* has evolved gradually in 2002–2021 through lessons learned when developing the *Sampo series* of semantic portals in use, including MuseumFinland (2004), CultureSampo (2009), BookSampo (2011), WarSampo (2015), BiographySampo (2018), NameSampo (2019), WarWictimSampo (2019), MMM (2020), AcademySampo (2021), and FindSampo (2021). These Semantic Web applications surveyed in this paper cover a wide range of application domains in CH and have attracted up to millions of users on the Semantic Web, suggesting feasibility of the proposed Sampo model. This work shows a shift of focus in research on CH semantic portals from data aggregation and exploration systems (1. generation systems) to systems supporting DH research (2. generation systems) with data analytic tools, and finally to automatic knowledge discovery and Artificial Intelligence (3. generation systems).

Keywords: Semantic Web, Digital Humanities, Linked open data, Data Services, Portals

1. Breaking Data Silos of Cultural Heritage

Cultural Heritage content is published independently by different memory organizations, such as museums, libraries, archives, galleries, and media companies. The traditional web publishing model, where everybody can publish easily content for everybody to read, facilitates fast and flexible publication on the Web. However, using related local contents from separate data sources on a global level is difficult because of the incompatible *data silos*: the local databases and online systems of the publishers are associated in content, but heterogeneous in terms of incompatible data models, annotated using different thesauri and vocabularies, distributed geographically, based on different natural languages, and used with different kind of user interfaces. An even more fundamental problem is that the contents are typically published only for humans to read and not as data for computational analyses and application development. This means that the end users typically have to learn and use several different applications to cater their information needs about a topic. For the data publishers, lots of costly redundant work is needed in creating the data silos, e.g., in developing the vocabularies and data services. The availability of the data in a usable open form is a prerequisite of the work for the application developers.

To mitigate these problems, various massive international data aggregation systems have been created, such as Europeana[1] in Europe and the Digital Public Library of America[2] in the U.S. There are lots of similar systems around on a national and regional level

[1] https://europeana.org
[2] https://dp.la/

(e.g., Deutsche Digitale Bibliothek[3] in Germany and K-samsök service in Sweden) and within various thematic communities[4] (e.g., AriadnePLUS[5] in archaeology). Similar data aggregation systems have also been created within single organizations that may already have lots of siloed but related databases around, like in the case of BBC in the U.K. [1]. There are lots of international and national standardization efforts for creating harmonized data models (e.g., Dublin Core[6], CIDOC CRM[7], and FRBRoo[8] [2]), shared thesauri for annotating contents (e.g., AAT, TGN, and ULAN vocabularies of the Getty Research Institute[9]), as well as generic frameworks, such as the Semantic Web standards of W3C[10].

This paper concerns using Semantic Web (SW) technologies [3] and Linked Open Data (LOD) publishing [4, 5] to address the data silo and data publishing problems above. A general model, called *Sampo Model*, is presented for the purpose. As empirical evidence of feasibility of applying the model in practise, the *Sampo series* of semantic portals is presented[11]. They have had millions of users on the Semantic Web. The fundamental idea of Linked Data is to create an interoperable interlinked Web of Data [4]. The novelty of the Sampo model lays in its attempt to formulate and generalize this idea into a set of re-usable design principles or guide lines for creating semantic portals, especially for Cultural Heritage applications and Digital Humanities research [6]. The Sampo model is a kind of consolidated approach for creating LOD services and semantic portals, something that the field of the Semantic Web is arguably still largely missing [7].

This paper is organized as follows. Section 2 presents the principles of the Sampo model. In section 3, a survey of Sampo systems is presented as a proof-of-concept, illustrating use cases of the model and how it has evolved in 2002–2021. In conclusion, related works are discussed, contributions of this paper are summarized, and directions for further research are outlined. This paper extends substantially the earlier short paper [8] about the Sampo model at the DHN 2020 conference.

Table 1
Sampo Model Principles P1–P6

P1.	Support collaborative data creation and publishing
P2.	Use a shared open ontology infrastructure
P3.	Support data analysis and knowledge discovery in addition to data exploration
P4.	Provide multiple perspectives to the same data
P5.	Standardize portal usage by a simple filter-analyze two-step cycle
P6.	Make clear distinction between the LOD service and the user interface (UI)

2. Sampo Model Principles

Sampo Model is an informal collection of principles for LOD publishing and designing semantic portals listed in Table 1, supported by an ontology and data infrastructure and software tools for user interface design and data publication. The model is called "Sampo" according to the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology[12] according to the most common interpretation of the concept. The principles P1–P6 of Table 1 are described and motivated in more detail in the following subsections, one after another.

P1. Support collaborative data creation and publishing The model is based on the idea of collaborative content creation. The data is aggregated from local data silos into a global service, based on a shared ontology and publishing infrastructure [5]. The local data are harmonized and enriched with each other by linking and reasoning, based on Semantic Web standards. In this model everybody can win, including the data publishers by enriched data and shared publishing infra, and the end users by richer global content and services. However, collaborative publishing also complicates the publication process, as more agreements are needed within the community.

This model addresses the problems of semantic data interoperability and distributed content creation at the same time. A shared semantic ontology infrastructure that includes shared metadata schemas and domain ontologies for population the data models are used for

[3] https://www.deutsche-digitale-bibliothek.de/?lang=en
[4] See https://pro.europeana.eu/page/aggregators for a list such systems.
[5] https://ariadne-infrastructure.eu/
[6] https://dublincore.org/
[7] https:/cidoc-crm.org
[8] https://www.ifla.org/node/10171
[9] https://www.getty.edu/research/tools/vocabularies/
[10] https://www.w3.org/standards/semanticweb/
[11] See https://seco.cs.aalto.fi/applications/sampo/ for a complete list of "Sampo portals", videos, and further information.
[12] https://en.wikipedia.org/wiki/Sampo

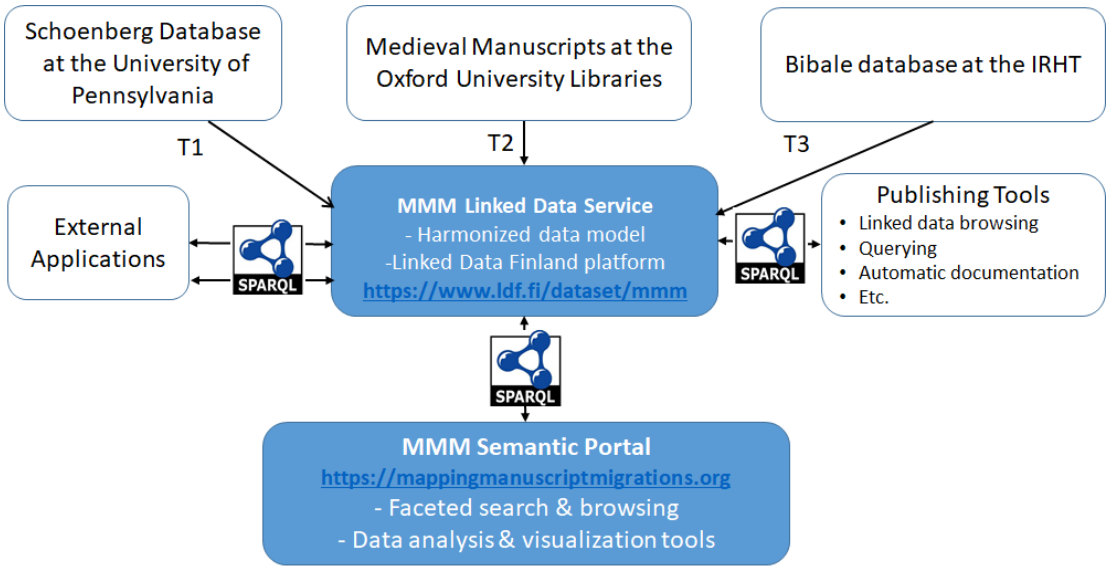

Fig. 1. Publishing and using heterogeneous distributed data in the MMM Sampo system

harmonizing and interlinking data form separate silos. If the content providers provide the system with metadata about their contents using the shared infrastructure, the data is automatically linked and enriched with each other and forms a knowledge graph [9]. 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.

Fig. 1 depicts as an example how the Sampo publication model (P1) was used in the Mapping Manuscript Migrations (MMM) system [10]. MMM includes three key datasets about ca. 220 000 medieval and Renaissance manuscripts that originate from the U.S. (Schoenberg Institute (T1)), U.K. (Oxford University Libraries (T2)), and France (the Institut de recherche et d'histoire des textes (IRHT) (T3)). The data T1–T3 are transformed into the unified harmonizing data model used in the MMM Linked Data Service [11] that is depicted in the middle of the figure. The data service is used by the MMM portal (bottom) but can also be used in other external applications via the SPARQL endpoint (on the left). The global data is documented and can be studied using SPARQL and publishing tools (on the right), too. The aggregated global data can be used for solving research questions that cannot be answered by studying the local datasets separately.

P2. Use a shared open ontology infrastructure The Sampo model is based on a shared LOD ontology infrastructure with which the local datasets are made compatible. Re-using the same infrastructure, and developing it further step by step in each Sampo portal and application, saves a lot of effort for the developers of next Sampos and other applications. For example, the linked data based geogazetteer of contemporary placenames in Finland, based on data from the National Survey and introduced in NameSampo [12] for open use, contains some 800 000 geocoded places, and there are other ontologies for historical places, maps, and persons.

The infrastructure includes harmonising shared metadata models (schemas) for representing individuas as well as domain ontologies (thesauri, vocabularies) that are used in populating (instantiating) the metadata models. This can be done by using data transformations and by aligning ontologies, as described in detail in [11, 13] for WarSampo and MMM systems, respectively. The Sampo portals use in practise both Dublin Core -based models and the dumb-down principle[13] for documents, and event-based models conforming

[13] https://dublincore.org/

and extending the CIDOC CRM ontology and FRBRoo. In addition to sharing same infrastructure components, different Sampos enrich each other's contents by mutual data linking, creating a gradually evolving network of Sampos, a kind of "SampoSampo" and data cloud. Also data from the international data infrastructure is used for this purpose, e.g., Wikidata[14] ja GeoNames[15]. The WarSampo knowledge graph [13], for instance, is part of the LOD Cloud[16].

The Sampo series systems that are based on Finnish datasets make use of the national FinnONTO ontology infrastructure [14]. Its development started in 2003 and is carried on today by the National Library of Finland as the Finto.fi ontology service[17], and under the research initiative Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)[18] [15].

P3. Support data analysis and knowledge discovery in addition to data exploration Three generations of semantic portals for Digital Humanities can be identified according to the vision [16] underlying the work on Sampos. Ten years ago the research focus in semantic portal development was on data harmonization, aggregation, search, and browsing. At the moment, the rise of DH research has started to shift the focus to providing the user with integrated tools for solving research problems in interactive ways. The next step ahead to is based on Artificial Intelligence: future portals not only provide tools for the human to solve problems but are used for finding research problems in the first place, for addressing them, and even for solving them automatically under the constraints set by the human researcher. Such systems should preferably be able to explain their reasoning, which is an important aspect in the source critical humanities research.

The Sampo mdel aims not only at data publishing with search and data exploration, as discussed, e.g., in [17], but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analysing, and even solving research problems in interactive ways, based on AI techniques [16].

P4. Provide multiple perspective to the same data THe Sampo model fosters the idea that on top of a LOD service different thematic *application perspectives* to the data can be created by re-using the data service. This means that the underlying data can be re-used without modifying it, which is typically costly [18] when dealing with Big Data.

The application perspectives are provided on the landing page of the Sampo portal system, and they enrich each other by data linking. By selecting a perspective the corresponding application is opened. In addition, completely separate applications can be created on top of the data service by third parties, which is of help to memory organizations that typically are not strong in IT application development but are often willing to share the content openly through multiple channels.

For example, Fig. 2 depicts the landing page of WarSampo [19] with the following nine interlinked application perspectives for accessing the underlying LOD service data:

1. Major events of WW2 visualized on a timeline and maps with related linked data
2. People (100 000) with biographical data and links to related perspectives
3. Army Units (15 900) including events, war diaries, and people related to the units
4. Places perspective for searching the war zone events using contemporary and historical maps
5. *Kansa taisteli* magazines (1957–1986) containing thousands of memoirs of the soldiers after the war
6. Casualties data (95 000 death records) of the soldiers killed in action
7. Authentic photographs (160 000) from the war zone by the Defence Forces of Finland, interlinked with people and places
8. War Cemeteries of the casualties in Finland (630) with 3000 photographs
9. Finnish Prisoners of War (4500) in the Soviet Union in 1939–1945

P5. Standardize portal usage by a simple filter-analyze two-step cycle In later Sampos, the application perspectives can be used by a two-step cycle for research: First the focus of interest, the target group, is filtered out using faceted semantic search [20–22]. Second, the target group is visualized or analyzed by using ready-to-use DH tools of the application perspectives. The general idea here is to try to "standardize" the UI logic so that the portals are easier to use for the end users [23].

For example, Fig. 3 depicts a situation in BiographySampo where the user compares the life charts of two prosopographical groups in 1809–1917 when Finland was an autonomous Grand Duchy within the Rus-

[14] https://www.wikidata.org/
[15] https://www.geonames.org/
[16] https://lod-cloud.net/
[17] https://finto.fi
[18] https://seco.cs.aalto.fi/projects/lodi4dh/

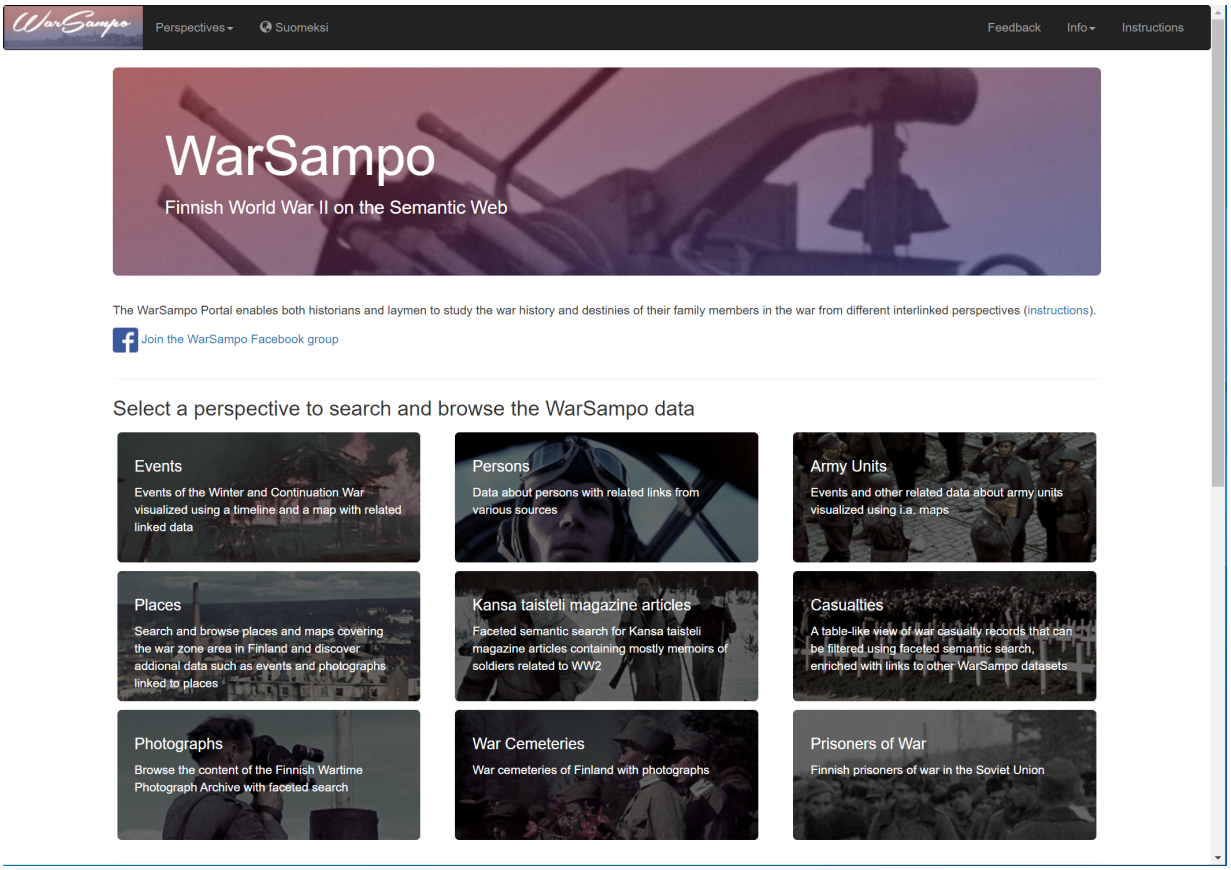

Fig. 2. Landing page of WarSampo with nine application perspectives

sian Empire: 1) Finnish generals and admirals in the Russian armed forces (on the left). 2) Members of the Finnish clergy (1800–1920) (on the right). With a few selections from the facets the user can filter out the two target groups and see that, for some reason, quite a few officers moved to Southern Europe when they retired, like retirees today, while the Lutheran ministers stayed in Finland.

P6. Make clear distinction between the LOD service and the user interface (UI) The Sampo Model argues for the idea of separating the underlying Linked Data service *completely* from the user interface via a SPARQL API. The rationale for this is: Firstly, this simplifies the portal architecture. Secondly, the data service can be opened for data analysis research in Digital Humanities. For example, YASGUI[19] [24] interface for SPARQL querying and visualizing the results can be used, or Python scripting in Google Colab[20] and Jupyter[21] [25].

The Sampo model principles above are compatible with the FAIR principles for creating Findable, Accessible, Interoperable, and Re-usable data[22], but were developed in the context of publishing and using Cultural Heritage Linked Open Data on the Semantic Web. The Sampo model can, however, be applied in other domains, too. An example of this is the HealthFinland system [26] for health promotion information, that was deployed by the National Institute for Health and Welfare in Finland[23].

[19] https://yasgui.triply.cc

[20] https://colab.research.google.com/notebooks/intro.ipynb

[21] https://jupyter.org

[22] https://www.go-fair.org/fair-principles/

[23] HealthFinland got at the ISWC 2008 conference the international Semantic Web Challenge Award.

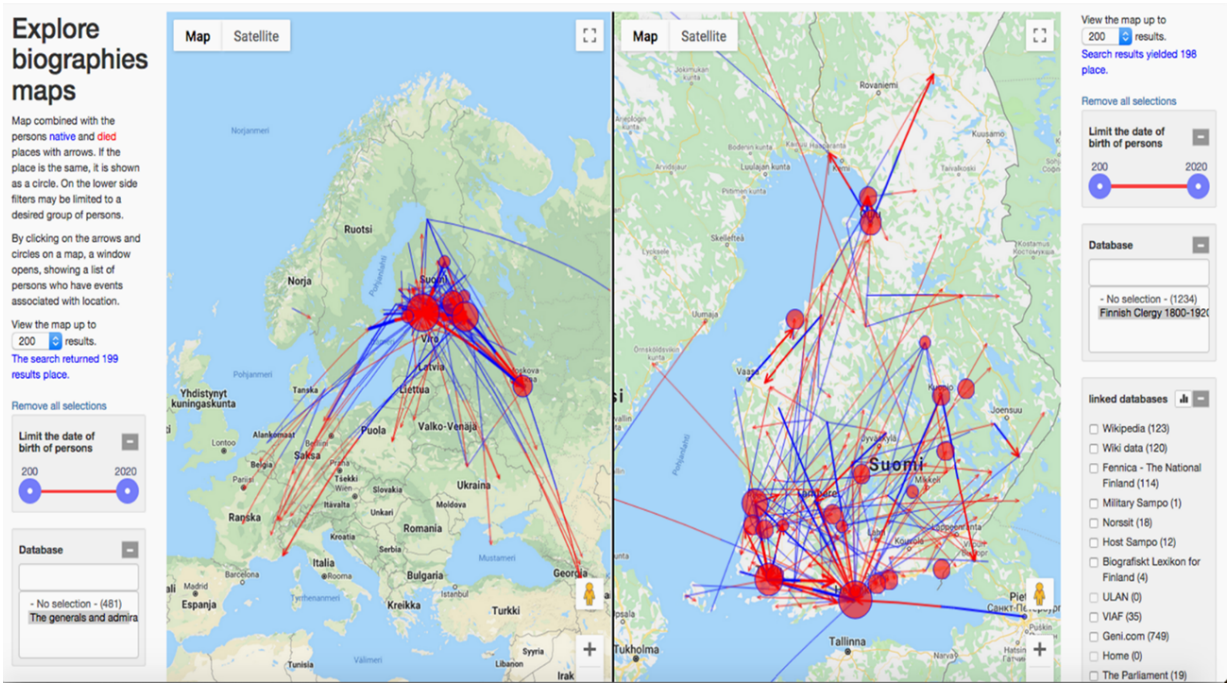

Fig. 3. Comparing the life charts of two target groups in BiographySampo, admirals and generals (left) and clergy (right) of the historical Grand Duchy of Finland (1809–1917).

3. Sampo Series of Semantic Portals and LOD Services

The Sampo model has evolved gradually over time in 2002–2021 via lessons learned in developing the Sampo series of semantic portals and related LOD services in various projects. This section overviews shortly a selection of these systems, listed in Table 3, in order to provide a proof-of-concept of the model and to give some examples and historical background of the work. For each system, the year of publication, application domain, number of end users, size of the underlying triplestore, and primary data owners are listed. In below, each system is described shortly with a reference to its research homepage and to at least one research article for more detailed information. These references provide links to related works and additional publications, and to the actual portals and web services online.

MuseumFinland – Finnish Museums on the Semantic Web[24] (online since 2004) [27] was the first Sampo. It introduced principle P1 of Table 1 by aggregating and publishing heterogeneous, distributed artefact collection data from Finnish museums. This application got the Semantic Web Challenge Award at the ISWC 2004 conference.

CultureSampo – Finnish Culture on the Semantic Web 2.0[25] (online since 2009) [18, 28] introduced principles P2 and P4. It demonstrated how CH content of tens of different kinds, both tangible and intangible CH content, can enrich each other. CultureSampo includes, e.g., a semantic model of the Kalevala epic narrative, based on a national ontology infrastructure. The name "Sampo" originates from this connection to the epic and has been re-used as a "brand" name in most of the offspring systems after that.

BookSampo – Finnish Fiction Literature on the Semantic Web[26] (online since 2011) [29] publishes metadata about virtually all Finnish fiction literature as a knowledge graph on top of which a portal was created. BookSampo data was originally part of CultureSampo but is today maintained independently by the Public Libraries of Finland. BookSampo has grown into one of the main web services of the Finnish public libraries, and is used by ca. 2 million users in a year.

[24] https://museosuomi.fi

[25] https://seco.cs.aalto.fi/applications/kulttuurisampo/

[26] https://seco.cs.aalto.fi/applications/kirjasampo/

Table 2
A selection of Sampo portals and LOD services for Digital Humanities

Portal	Year	Domain	# Users	# Triples	Primary data owners
MuseumFinland	2004	Artefact collections	39 000	211 000	National Museums, City Museums of Espoo and Lahti, Finland
CultureSampo	2008	Finnish culture	107 000	11M	Memory organizations and the Web, ca 30 data sources
BookSampo	2011–	Fiction literature	2M/year	4,36M[a]	Public libraries in Finland (Kirjastot.fi)
WarSampo	2015–2019	World War II	740 000	14M	National Archives, Defense Forces, and others, Finland
Norssit Alumni	2017	Person registry	unknown	469 000	Norssi High School alumni organization Vanhat Norssit
U.S. Legislator Prosographer	2018	Parliamentary data	unknown	830 000	U. S. Congress Legislator data[b]
NameSampo	2019	Place names	35 000	26,0M[c]	Institute for the Languages of Finland (Kotus), National Land Survey of Finland, and the J. Paul Getty Trust TGN Thesaurus
BiographySampo	2019	Biographies	50 000	5,56M	Finnish Literature Society
WarVictimSampo 1914–1922	2019	Military history	29 000	9,96M	National Archives of Finland
Mapping Manuscript Migrations (MMM)	2020	Pre-modern manuscripts	2200	22,5M	Schoenberg Inst. for Manuscript Studies (U.S.), Oxford University Libraries (U.K.), and Inst. for Research and History of Texts (France)
AcademySampo	2021	Finnish Academics	2100	6,55M	University of Helsinki and National Archives, Finland
FindSampo	2021	Archaeology, finds	1100	1,0M	Finnish Heritage Agency, Finland

[a]Original KG size in 2011; the size in much larger today including also non-fiction works
[b]https://github.com/unitedstates/congress-legislators
[c]This count includes only data of Kotus; the total number of triples of all sources is 241M.

WarSampo – Finnish World War II on the Semantic Web[27] (online since 2015 with several new perspectives published in 2016–2019) [19] is a popular Finnish service that has had 740 000 users. It introduced principle P6 into the Sampo model. The portal and its data service provides information about the casualties and significant soldiers of the Second World War in Finland. The dataset includes various graphs, such as authentic photographs from the fronts, war diaries, historical maps, memoir articles of soldiers, etc., constituting small a LOD cloud of its own and an infrastructure for Finnish WW2 data [13]. WarSampo application got in 2017 the LODLAM Open Data Prize in Venice.

Interest in WarSampo lead to a new Sampo in the same application domain of war history: **WarVictimSampo (1914–1922)**[28] (online since 2019) [30] publishes data about the deaths and battles of the Finnish Civil War 1918 and related wars. Also this portal has become fairly popular, as many citizens are still looking for information about their lost relatives in the Civil War. Both WarVictimSampo and WarSampo have a feedback channel by which the data can be commented, and indeed hundreds of comments and suggestions for corrections have been collected for the data owner, the National Archives of Finland, to consider. Based on this activity, a new citizen science project for collecting and maintaining Sampo data is currently underway[29].

A key idea in WarSampo is to reassemble the life stories of the soldiers based on data linking from different data sources. This biographical and prosopo-

[27]https://seco.cs.aalto.fi/projects/sotasampo/en/

[28]https://seco.cs.aalto.fi/projects/sotasurmat/

[29]https://seco.cs.aalto.fi/projects/sotasampo/citizens/en/

graphical idea was a source of inspiration for several later biographical Sampos discussed below.

BiographySampo – Biographies on the Semantic Web[30] (online since 2018) [31] is yet another popular service with tens of thousands of users. It harnessed principles P3 and P5 into the Sampo model, with a focus on supporting biographical and prosopographical research and data analysis. The system is based on mining out a large knowledge graph from ca. 13 100 Finnish national biographies of the Finnish Literature Society, authored by some 940 scholars. The data is interlinked and enriched internally and by 16 external data sources and by reasoning, e.g., family relations [32] and serendipitous connections between people and places [33].

The idea of publishing textual biographies as structured LOD for data exploration and analysis was also developed in the Sampos **Norssit Alumni** [34] and **U.S. Congress Prosopographer** [35]. **AcademySampo**[31] (online since 2021) [32] is yet another biographical system based on 28 000 short biographies of all known Finnish academic people educated in Finland in 1640–1899.

NameSampo – A Linked Open Data Infrastructure and Workbench for Toponomastic Research[32] (online since 2019) [12] publishes data about over 2 million place names and places in Finland with old maps. It soon attracted tens of thousands of users on the Web. NameSampo core data originates from the Name Archive of the Institute of Languages of Finland, a database of over 2 million placenames collected in Finland over several decades. NameSampo also published the contemporary placename register (ca. 800 000 places) of the National Survey of Finland as Linked Open Data. Furthermore, the Thesaurus of Geographical Names (TGN)[33] of Getty Research via its SPARQL endpoint is re-used, as well as various map services, including a collection of historical maps of Finland published as part of WarSampo.

The NameSampo project developed, based on the SPARQL Faceter tool [36] used in many earlier Sampos, the first version of the Sampo-UI framework [23] that has been used after this is all Sampos, supporting implementation of principles P3–P5 from an UI point of view. Sampo-UI has also been reused in Norway by the Norwegian Language Collections for creating a national service similar to NameSampo: Norske stedsnavn[34]. The Sampo-UI framework, available openly in Github[35], has also been re-used in a commercial setting.

Mapping Manuscript Migrations (MMM)[36] (online since 2020) [10, 11] is a Sampo, in spite of it name, based on metadata about some 220 000 premodern manuscripts from the Schoenberg Database of Manuscripts[37] in the U.S., Medieval Manuscripts in Oxford University Libraries[38] in the U.K., and the Bibale[39] database in France. MMM is a result of the Trans-Atlantic Digging into Data research programme[40].

FindSampo[41] [37] (online since 2021) is a system and data service for supporting archaeology especially form a citizen science and metal detectorists' perspectives.

In addition, new Sampos are already in prototype phase: **LawSampo**[42] [38] publishes Finnish legislation and case law based on data from the Ministry of Justice in Finland. **ParliamentSampo**[43] publishes LOD extracted from the materials of the Parliament of Finland (1907–2021)[44], including over 900 000 Parliamentary debate speeches [39] and prosopographical data about the politicians' networks [40] in 1907–2021. **LetterSampo**[45] [41] is based on early modern epistolary metadata aggregated in the Early Modern Letters Online (EMLO) service[46] at the Oxford University, the CKCC corpus underlying ePistolarium[47] of the Huygens Institute in the Netherlands, and correspSearch[48] service of the Berlin-Brandenburg Academy of Sciences.

[30] https://seco.cs.aalto.fi/projects/biografiasampo/en/
[31] https://seco.cs.aalto.fi/projects/akatemiasampo/en/
[32] https://seco.cs.aalto.fi/projects/nimisampo/en/
[33] http://www.getty.edu/research/tools/vocabularies/tgn/
[34] https://toponymi.spraksamlingane.no/nb/app
[35] https://github.com/SemanticComputing/sampo-ui
[36] https://seco.cs.aalto.fi/projects/mmm/
[37] See https://sdbm.library.upenn.edu
[38] See https://medieval.bodleian.ox.ac.uk
[39] The Bibale web service from the Institute for Research and History of Texts (IRHT) in Paris is described in http://bibale.irht.cnrs.fr.
[40] https://diggingintodata.org/
[41] https://seco.cs.aalto.fi/projects/sualt/
[42] https://seco.cs.aalto.fi/projects/lawlod/
[43] https://seco.cs.aalto.fi/projects/semparl/en/
[44] https://seco.cs.aalto.fi/projects/semparl/en/
[45] https://seco.cs.aalto.fi/projects/rrl/
[46] http://emlo.bodleian.ox.ac.uk
[47] http://ckcc.huygens.knaw.nl/epistolarium/
[48] https://correspsearch.net

4. Discussion

The idea of trying to formulate general design principles for publishing and using linked data has turned out to be useful from a practical point of view. For example, the four Linked Data Principles[49] and the 5-star model[50] coined by Tim Berners-Lee have been quite influential, and ontology design patterns[51] are (re-)used as guidelines for data modelling. In the same vein, the FAIR principles for publishing data are widely used today. Design principles, models, and methods for software development are extensively studied and used in the field of Software Engineering [42]. From these perspectives, the Sampo model can be seen as a kind of hybrid effort for formulating a set of principles for publishing and using linked data in semantic portals, especially for Digital Humanities. Our experiences in developing the Sampo series of data services and portals in 2002–2021 provide an empirical evaluation or evidence about the usability of the model in practical applications. The application domains of the model (cf. Table 3) are versatile including tangible and intangible cultural heritage collections, biography and prosopography, toponomastic research, manuscript studies, archaeology, legislation, and parliamentary studies. In many cases, language barriers have been crossed based on the language-agnostic ontology technology [43] of the Semantic Web.

Related Work The Principles of Table 1 behind the Sampo model have been explored and developed before in different contexts:

1. The principle of collaborative content creation by data linking (P1) is a fundamental idea behind the Linked Open Data Cloud movement[52] and has been developed also in various other settings, e.g., in ResearchSpace[53].
2. The importance of developing shared open data models, thesauri, and ontologies for interoperability (P2) is a driving force behind the work of virtually all related standardization efforts. In our work, the ambitious goal has been to develop not only individual standards and datasets but an infrastructure on a national level effort [14] in terms of open ontology services [44, 45] and LOD services [46].
3. The principle of supporting data analysis and knowledge discovery (P3) based on Big Data is fundamental in, e.g., distant reading [47], Humanities Computing [48], and Digital Humanities [6] in general. However, what is still largely missing in the DH methodology and tools in semantic portals is the next conceptual level of automatic knowledge discovery from data [49]. The Sampo model aims to integrate such tools into a consolidated approach for creating portals and LOD services.
4. The principle P4 of providing multiple analyses and visualizations for a set of filtered search results has been used in different contexts and also in other portals, such as the ePistolarium[54] [50] for epistolary data. The idea of using multiple perspectives has also been studied as an approach in decision making [51].
5. Regarding principle P5, faceted search [20, 21, 52], also know as "view-based search" and "dynamic ontologies", is a well-known paradigm for explorative search and browsing [17] in computer science and information retrieval, based on S. R. Ranaganathan's original ideas of faceted classification in Libary Science. The two step usage model in Sampo model is also used as a general research method in prosopographical research [53].
6. The principle P6 of separating data related services from UI design is in line with modern software architectures, such as the Model-View-Controller (MVC) structure[55]. The Sampo model supports the idea of "separation of concerns" where each software layer can focus solely on its own role, and uses the Web idea of using the simple HTTP protocol for creating services based on other distributed services.

Contributions and Challenges The novelty of the Sampo model lies in the consolidated combination of the principles P1–P6 and in operationalizing them using an infrastructure and tooling for developing applications in Digital Humanities in a cost-efficient way. The approach aims at developing a gradually growing sustainable national LOD infrastructure: the work started with the Semantic Web Kick-off in Finland seminar [54] a few months after the seminal Semantic Web paper [55] was published in Scientific Amer-

[49] https://www.w3.org/DesignIssues/LinkedData
[50] https://5stardata.info/en/
[51] http://ontologydesignpatterns.org/wiki/Main_Page
[52] https://lod-cloud.net
[53] https://www.researchspace.org
[54] http://ckcc.huygens.knaw.nl
[55] https://en.wikipedia.org/wiki/Model-view-controller

ican and W3C launched its Semantic Web Activity in 2001. The work presented demonstrates a shift of focus in research on CH semantic portals from data aggregation and exploration systems (1. generation systems) to systems supporting DH research (2. generation systems) with data analytic tools, and finally to automatic knowledge discovery and Artificial Intelligence (3. generation systems) [16].

The model has also its limitations and challenges. For example, it does not include any principles for maintaining the knowledge graphs but assumes that the data is created by a separate pipeline. As suggested in [13], the transformation should be automatic and re-doable without a human in the loop, but optimally the RDF should be produced already when cataloging the data, not by correcting and aligning the data afterwards. As Alfred Einstein put it: "Intellectuals solve problems but geniuses prevent them".

A major challenge of the semantic portal concept is related to the quality of the data produced typically using more or less automatic means, leading to problems of incomplete, skewed, and erroneous data. This as well as conceptual difficulties in modeling complex real world ontologies, such as historical geogazetteers, become sometimes embarrassingly visible when using and exposing the knowledge structures to end-users. In traditional systems the same problems are there, but are hidden in the non-structured presentations of the data. In general, more data literacy [56] is usually needed from the end-user when using semantic portals and their data analytic tools. In spite of these challenges the linked data approach is according to our experiences useful is finding out interesting phenomena in Big Data using distant reading [47], but for interpreting the results also traditional close reading is needed as before.

Future Research The future work on Sampo model aims at AI based DH tools that are able not only to present the data to the human researcher in useful ways but also to 1) find DH research problems, 2) solve them *automatically by themselves*, and 3) also explain the reasoning or solution to the researcher. AI techniques would also be useful when creating and enriching the knowledge graph underlying a semantic portal. First steps towards these goals have already been taken, e.g., in BiographySampo where the underlying knowledge graph has been used for discovering and explaining serendipitous semantic connections between places and persons to the end user [31, 33].

Acknowledgments Tens of people[56] at the Semantic Computing Research Group (SeCo) have been working in developing the Sampo model, portals, and infrastructure, funded by over 50 organizations in 2002–2021 in Finland and beyond. The work of this paper is partly funded by the Semantic Parliament (ParliamentSampo) project of the Academy of Finland, the EU project InTaVia: In/Tangible European Heritage[57], and the EU COST action Nexus Linguarum[58] on linguistic data science.

References

[1] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer and R. Lee, Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections, in: *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 723–737. ISBN 978-3-642-02121-3.

[2] P. Riva, M. Doerr and M. Zumer, FRBRoo: Enabling a Common View of Information from Memory Institutions, *International Cataloguing and Bibliographic Control (ICBC)* **38**(2) (2009).

[3] P. Hitzler, M. Krötzsch and S. Rudolph, *Foundations of Semantic Web technologies*, Springer, 2010.

[4] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space (1st edition)*, Morgan & Claypool, Palo Alto, California, 2011. http://linkeddatabook.com/editions/1.0/.

[5] E. Hyvönen, *Publishing and using cultural heritage linked data on the Semantic Web*, Morgan & Claypool, Palo Alto, California, 2012.

[6] E. Gardiner and R.G. Musto, *The Digital Humanities: A Primer for Students and Scholars*, Cambridge University Press, New York, NY, USA, 2015, https://doi.org/10.1017/CBO9781139003865.

[7] P. Hitzler, A Review of the Semantic Web Field, *Commun. ACM* **64**(2) (2021), 76–83–. doi:10.1145/3397512.

[8] E. Hyvönen, "Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web, in: *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, vol. 2612, 2020, pp. 373–378. http://ceur-ws.org/Vol-2612/poster1.pdf.

[9] C. Gutierrez and J.F. Sequeda, Knowledge graphs, *Communications of the ACM* **64**(3) (2021), 96–104. doi:10.1145/3418294.

[10] E. Hyvönen, E. Ikkala, J. Tuominen, M. Koho, T. Burrows, L. Ransom and H. Wijsman, A Linked Open Data Service and Portal for Pre-modern Manuscript Research, in: *DHN 2019 Digital Humanities in Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, CEUR Workshop Proceedings, Vol-2364, 2019, pp. 220–229.

[56] https://seco.cs.aalto.fi/people/

[57] https://intavia.eu/

[58] https://nexuslinguarum.eu/the-action

[11] M. Koho, T. Burrows, E. Hyvönen, E. Ikkala, K. Page, L. Ransom, J. Tuominen, D. Emery, M. Fraas, B. Heller, D. Lewis, A. Morrison, G. Porte, E. Thomson, A. Velios and H. Wijsman, Harmonizing and Publishing Heterogeneous Pre-Modern Manuscript Metadata as Linked Open Data, 2021, Accepted, JASIST Special Issue.

[12] E. Ikkala, J. Tuominen, J. Raunamaa, T. Aalto, T. Ainiala, H. Uusitalo and E. Hyvönen, NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research, in: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'18, ACM, New York, NY, USA, 2018, pp. 2:1–2:9. ISBN 978-1-4503-6032-6. doi:10.1145/3282933.3282936.

[13] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen and E. Hyvönen, WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data, *Semantic Web* **12**(2) (2021), 265–278. doi:10.3233/SW-200392.

[14] E. Hyvönen, K. Viljanen, J. Tuominen and K. Seppälä, Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach, in: *Proceedings of the ESWC 2008, Tenerife, Spain*, Springer, 2008, pp. 95–109.

[15] E. Hyvönen, Linked Open Data Infrastructure for Digital Humanities in Finland, in: *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, vol. 2612, 2020, pp. 254–259. http://ceur-ws.org/Vol-2612/short10.pdf.

[16] E. Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, *Semantic Web* **11**(1) (2020), 187–193.

[17] G. Marchionini, Exploratory search: from finding to understanding, *Communications of the ACM* **49**(4) (2006), 41–46.

[18] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio and K. Nyberg, CultureSampo – Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user, in: *Museums and the Web 2009*, Archives & Museum Informatics, Toronto, 2009.

[19] E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen and E. Mäkelä, 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)*, Springer, 2016, pp. 758–773.

[20] E. Hyvönen, S. Saarela and K. Viljanen, Application of ontology-based techniques to view-based semantic search and browsing, in: *Proceedings of the First European Semantic Web Symposium*, Springer, 2004.

[21] D. Tunkelang, *Faceted search*, Morgan & Claypool Publishers, CA, USA, 2009.

[22] Y. Tzitzikas, N. Manolis and P. Papadakos, Faceted exploration of RDF/S datasets: a survey, *Journal of Intelligent Information Systems* **48**(2) (2017), 329–364.

[23] E. Ikkala, E. Hyvönen, H. Rantala and M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, *Semantic Web* (2021), Accepted.

[24] L. Rietveld and R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web* **8**(3) (2017), 373–383. doi:10.3233/SW-150197.

[25] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen and E. Hyvönen, Automatic Annotation Service: Utilizing a Named Entity Linking Tool in Legal Domain, 2019, Submitted article under evaluation.

[26] O. Suominen, E. Hyvönen, K. Viljanen and E. Hukka, HealthFinland – a National Semantic Publishing Network and Portal for Health Information, *Journal of Web Semantics* **7**(4) (2009), 287–297.

[27] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila and S. Kettula, MuseumFinland—Finnish Museums on the Semantic Web, *Journal of Web Semantics* **3**(2) (2005), 224–241.

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

[29] E. Mäkelä, K. Hypén and E. Hyvönen, BookSampo—Lessons Learned in Creating a Semantic Portal for Fiction Literature, in: *Proc. of ISWC-2011, Bonn, Germany*, Springer, 2011.

[30] H. Rantala, E. Ikkala, I. Jokipii, M. Koho, J. Tuominen and E. Hyvönen, WarVictimSampo 1914–1922: a National War Memorial on the Semantic Web for Digital Humanities Research and Applications, *ACM Journal on Computing and Cultural Heritage* (2021), Accepted.

[31] 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 Extended Semantic Web Conference (ESWC 2019)*, Springer, 2019, pp. 574–589.

[32] P. Leskinen and E. Hyvönen, Linked Open Data Service about Historical Finnish Academic People in 1640–1899, in: *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, vol. 2612, 2020, pp. 284–292. http://ceur-ws.org/Vol-2612/short14.pdf.

[33] E. Hyvönen and H. Rantala, Knowledge-based Relational Search in Cultural Heritage Linked Data, *Digital Scholarship in the Humanities (DSH), Oxford University Press* (2021), In press.

[34] E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen and L. Sirola, Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web, in: *Proceedings, Language, Technology and Knowledge (LDK 2017)*, Springer, 2017, pp. 113–119. https://link.springer.com/chapter/10.1007/978-3-319-59888-8_9.

[35] G. Miyakita, P. Leskinen and E. Hyvönen, Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators, in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus*, Springer, 2018.

[36] M. Koho, E. Heino and E. Hyvönen, 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*, CEUR Workshop Proceedings, Vol. 1615, 2016. http://ceur-ws.org/Vol-1615/semdevPaper5.pdf.

[37] E. Hyvönen, H. Rantala, E. Ikkala, M. Koho, J. Tuominen, B. Anafi, S. Thomas, A. Wessman, E. Oksanen, V. Rohiola, J. Kuitunen and M. Ryyppö, Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework, *Antiquity, A Review of World Archaeology* (2021), In press.

[38] E. Hyvönen, M. Tamper, E. Ikkala, S. Sarsa, A. Oksanen, J. Tuominen and A. Hietanen, Publishing and Using Legislation and Case Law as Linked Open Data on the Semantic Web, in: *The Semantic Web: ESWC 2020 Satellite Events*, Lecture Notes in Computer Science, Vol. 12124, Springer, 2020, pp. 110–114. doi:10.1007/978-3-030-62327-2_19.

[39] L. Sinikallio, S. Drobac, M. Tamper, R. Leal, M. Koho, J. Tuominen, M.L. Mela and E. Hyvönen, Plenary debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN markup, in: *Proceedings, Language, Data and Knowledge (LDK 2021)*, 2021, In press.

[40] P. Leskinen, E. Hyvönen and J. Tuominen, Members of Parliament in Finland Knowledge Graph and Its Linked Open Data Service, in: *Proceedings of SEMANTiCS - In the Era of Knowledge Graphs, Amsterdam, Sept 6-9, 2021*, Studies on the Semantic Web, IOS Press, 2021, Accepted.

[41] J. Tuominen, E. Mäkelä, E. Hyvönen, A. Bosse, M. Lewis and H. Hotson, Reassembling the Republic of Letters – A Linked Data Approach, in: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, CEUR Workshop Proceedings, vol. 2084, 2018, pp. 76–88. http://www.ceur-ws.org/Vol-2084/paper6.pdf.

[42] I. Sommerville, *Software engineering (10th edition)*, Pearson, 2016.

[43] S. Staab and R. Studer (eds), *Handbook on Ontologies (2nd Edition)*, Springer, 2009.

[44] J. Tuominen, M. Frosterus, K. Viljanen and E. Hyvönen, ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services, in: *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, Springer, 2009.

[45] K. Viljanen, J. Tuominen and E. Hyvönen, Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI, in: *Proceedings of the ESWC 2009, Heraklion, Greece*, Springer, 2009, pp. 781–795.

[46] E. Hyvönen, J. Tuominen, M. Alonen and E. Mäkelä, Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets, in: *ESWC 2014: The Semantic Web: ESWC 2014 Satellite Events*, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

[47] F. Moretti, *Distant Reading*, Verso Books, 2013.

[48] W. McCarty, *Humanities Computing*, Palgrave, London, 2005.

[49] M.J. Pazzani, Knowledge discovery from data?, *IEEE Intelligent Systems* **15** (2000), 10–13.

[50] W. Ravenek, C. van den Heuvel and G. Gerritsen, The ePistolarium: Origins and Techniques, in: *CLARIN in the Low Countries*, A. van Hessen and J. Odijk, eds, Ubiquity Press, 2017, pp. 317–323. doi:10.5334/bbi.

[51] H.A. Linstone, Multiple perspectives: Concept, applications, and user guidelines, *Systems practice* **2**(3) (1989), 307–331. doi:10.1007/BF01059977.

[52] M. Hearst, Design recommendations for hierarchical faceted search interfaces, in: *ACM SIGIR workshop on faceted search*, Seattle, WA, 2006, pp. 1–5.

[53] 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. doi:1854/8212.

[54] E. Hyvönen (ed.), Semantic Web Kick-Off in Finland – Vision, Technologies, Research, and Applications, in *HIIT Publications 2002-01*, 2002. http://www.seco.hut.fi/publications/2002/hyvonen-semantic-web-kick-off-2002.pdf.

[55] T. Berners-Lee, J. Hendler and O. Lassila, The Semantic Web, *Scientific American* **284**(5) (2001), 34–43.

[56] T. Koltay, Data literacy for researchers and data librarians, *Journal of Librarianship and Information Science* **49**(1) (2015), 3–14. doi:10.1177/0961000615616450.