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

Eero Hyvönen[0000−0003−1695−5840]

Aalto University and University of Helsinki  
Helsinki Centre for Digital Humanities (HELDIG)  
Semantic Computing Research Group (SeCo)  
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 paper addresses these problems by presenting a collaborative publication model for CH Linked Data and an approach for creating shared data services and semantic portals for DH research and applications. This “Sampo Model” is based on a set on design principles, a shared ontology infrastructure and tools, and the idea of integrating data-analytic tools with traditional search and exploration of the data. The model has evolved gradually in 2002–2021 through lessons learned when developing the “Sampo series” of proof-of-concept CH systems, including MuseumFinland (2004), CultureSampo (2009), BookSampo (2011), WarSampo (2015), BiographySampo (2018), NameSampo (2019), WarVictimSampo (2019), MMM (2020), and AcademySampo (2021). These online systems 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 · Portals · Data Services · Digital Humanities

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 the 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 semantically interlinked 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 hu-
mans 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 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 re-
gegional level (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 archaeol-
ogy). 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. \(^{[26]}\). There are lots of international and national standardization
efforts for creating harmonized data models (e.g., Dublin Core, CIDOC CRM, and FR-
BRoo), shared thesauri for annotating contents (e.g., AAT, TGN, and ULAN of Getty
Research), as well as generic frameworks, such as the Semantic Web standards of W3C.

This paper argues for using Semantic Web (SW) technologies \(^6\) and Linked Open
Data (LOD) publishing \(^4,10\) to address the data silo and data publishing problems
above, and presents a general model, called \textit{Sampo Model} for the purpose. The idea of
using LOD for data aggregation is not new, but based on the fundamental idea of the
Semantic Web as an interoperable interlinked Web of Data. The novelty of the Sampo
model lays in its attempt for formulate and generalize the idea into a set re-usable prin-
ciples or guide lines, software tools, and ontology infrastructure, instead of just creating
individual LOD applications or tools. 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 \(^5\).

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 time 2002–2021. In con-
clusion, related works are discussed, contributions of this paper are summarized, and
directions for further research are outlined. This paper extends substantially the earlier
poster paper \(^{[14]}\) about the Sampo model at the DHN 2020 conference.

\section{Sampo Model Principles}

\textbf{Sampo Model} is an informal collection of principles for LOD publishing and design-
ing semantic portals listed in Table 1, supported by an ontology and data infrastructure

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\textbf{Principle} & \textbf{Description} \\
\hline
1 & https://europeana.org \\
2 & https://dp.la/ \\
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/
\hline
\end{tabular}
\end{table}
Table 1. Sampo Model Principles P1–P6

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

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 according to the most common interpretation of the concept.

Fig. 1. Publishing heterogeneous distributed data on the Semantic Web [10]

The principles P1–P6 of Table 1 are described and motivated in more detail in the following subsections, one after another.

**P1. Collaborative Data Creation and Publishing Model** 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 [10]. 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.

---


7 [https://www.w3.org/standards/semanticweb/](https://www.w3.org/standards/semanticweb/)
This model addresses 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 shared domain ontologies and metadata schemas, modeled by using SW standards. If content providers outside of the circle provide the system with metadata about their contents, the data is automatically linked and enriched with each other and forms a knowledge graph represented using RDF\(^8\). 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.

P2. Using a Shared Ontology Infrastructure As illustrated in Fig. 1, 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 [25] 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 individuals 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 [28,27] for WarSampo and MMM systems, respectively. The Sampo portals use in practise both Dublin Core-based models and the dumb-down principle\(^9\) for documents, and event-based models conforming to the CIDOC CRM ontology\(^10\) and FRBRO\(^11\). In addition to sharing same infrastructure components, 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 ja GeoNames. WarSampo, for instance, is part of the LOD Cloud\(^12\).

The Sampo series is based on the national FinnONTO ontology infrastructure [9], whose development that started 2003 and is carried on today by the National Library of Finland as the Finto.fi ontology service\(^13\), and under the research initiative Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)\(^14\) [13].

P3. From Data Publishing to Data Analysis and Knowledge Discovery Three generations of semantic portals for Digital Humanities can be identified according to

---

8 https://www.w3.org/RDF/
9 https://dublincore.org/
10 http://www.cidoc-crm.org/
11 https://www.ifla.org/node/10171
12 https://lod-cloud.net/
13 https://finto.fi
14 https://seco.cs.aalto.fi/projects/lodi4dh/
the vision [15] underlying the work on Sampo. Ten years ago the research focus in semantic portal development was on data harmonization, aggregation, search, and browsing (“first generation systems”). 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 (“second generation systems”). The next step ahead to “third generation systems” 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 Model aims not only at data publishing with search and data exploration, as discussed, e.g., in [35], 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.

P4. Multiple Perspective Interface Design

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 [11] 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 [16] 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 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. Photographs (160 000) from the war zone by the Defence Forces interlinked with persons and places
8. War Cemeteries of the casualties above in Finland (630) with 3000 photographs
9. Finnish Prisoners of War (4500) in the Soviet Union in 1939–1945

P5. Filter-analyze Two-step Usage 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 [8,45,48]. 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 [24].

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 Russian 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. Data Service – User Interface Distinction** 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\(^{15}\) [42] interface for SPARQL querying and visualizing the results can be used, or Python scripting in Google Colab\(^{16}\) and Jupyter\(^{17}\).

The Sampo model principles above are compatible with the FAIR principles for creating Findable, Accessible, Interoperable, and Re-usable data\(^{18}\), but developed further 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

---

\(^{15}\) [https://yasgui.triply.cc](https://yasgui.triply.cc)

\(^{16}\) [https://colab.research.google.com/notebooks/intro.ipynb](https://colab.research.google.com/notebooks/intro.ipynb)

\(^{17}\) [https://jupyter.org](https://jupyter.org)

\(^{18}\) [https://www.go-fair.org/fair-principles/](https://www.go-fair.org/fair-principles/)
example of this is the HealthFinland system [44] for health promotion information, that was deployed by the National Institute for Health and Welfare in Finland ten years ago.

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”19 of semantic portals and related LOD services in research 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 it. 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 Web20 (online since 2004) [7] was the first Sampo. It introduced Principle (1) of Table 1 by aggregating and publishing heterogeneous, distributed artefact collection data from Finnish museums. The application got the Semantic Web Challenge Award at the ISWC 2004 conference.

19 See https://seco.cs.aalto.fi/applications/sampo/ for a complete list of “Sampo portals”, videos, and further information.
20 https://museosuomi.fi
Table 2. A selection of Sampo portals and LOD services for Digital Humanities

<table>
<thead>
<tr>
<th>Portal</th>
<th>Year</th>
<th>Domain</th>
<th># Users</th>
<th># Triples</th>
<th>Primary data owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>MuseumFinland</td>
<td>2004</td>
<td>Artefact collections</td>
<td>39 000</td>
<td>211 000</td>
<td>National Museums, City Museums of Espoo and Lahti, Finland</td>
</tr>
<tr>
<td>CultureSampo</td>
<td>2008</td>
<td>Finnish culture</td>
<td>107 000</td>
<td>11M</td>
<td>Memory organizations and the Web, ca 30 data sources</td>
</tr>
<tr>
<td>BookSampo</td>
<td>2011–</td>
<td>Fiction literature</td>
<td>2M/year</td>
<td>4,36M(^a)</td>
<td>Public libraries in Finland (Kirjastot.fi)</td>
</tr>
<tr>
<td>WarSampo</td>
<td>2015–19</td>
<td>World War II</td>
<td>740 000</td>
<td>14M</td>
<td>National Archives, Defense Forces, and others, Finland</td>
</tr>
<tr>
<td>Norssit Alumni</td>
<td>2017</td>
<td>Person registry</td>
<td>unknown</td>
<td>409 000</td>
<td>Norssi High School alumni organization Vanhat Norssit</td>
</tr>
<tr>
<td>U.S. Legislator Prosographer</td>
<td>2018</td>
<td>Parliamentary data</td>
<td>unknown</td>
<td>830 000</td>
<td>U. S. Congress Legislator data(^b)</td>
</tr>
<tr>
<td>NameSampo</td>
<td>2019</td>
<td>Place names</td>
<td>35 000</td>
<td>26,0M(^c)</td>
<td>Institute for the Languages of Finland (Kotus), National Land Survey of Finland, and the J. Paul Getty Trust TGN Thesaurus</td>
</tr>
<tr>
<td>BiographySampo</td>
<td>2019</td>
<td>Biographies</td>
<td>50 000</td>
<td>5,56M</td>
<td>Finnish Literature Society</td>
</tr>
<tr>
<td>WarVictimSampo</td>
<td>1914–1922</td>
<td>Military history</td>
<td>29 000</td>
<td>9,96M</td>
<td>National Archives of Finland</td>
</tr>
<tr>
<td>Mapping Manuscript Migrations</td>
<td>2020</td>
<td>Pre-modern manuscripts</td>
<td>2200</td>
<td>22,5M</td>
<td>Schoenberg Inst. for Manuscript Studies (U.S.), Oxford Libraries (Oxford), and Inst. for Research and History of Texts (France)</td>
</tr>
<tr>
<td>AcademySampo</td>
<td>2021</td>
<td>Finnish Academics</td>
<td>2100</td>
<td>6,55M</td>
<td>University of Helsinki and National Archives, Finland</td>
</tr>
</tbody>
</table>

\(^a\) Original KG size in 2011; the size is 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.

**CultureSampo – Finnish Culture on the Semantic Web 2.0**\(^{21}\) (online since 2009) [11,34] introduced Principles (2) and (4). It demonstrated how CH content of tens of different kinds, both tangible and intangible CH content, can enrich each other, including 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**\(^{22}\) (online since 2011) [33] 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.

\(^{21}\) https://seco.cs.aalto.fi/applications/kulttuurisampo/
\(^{22}\) https://seco.cs.aalto.fi/applications/kirjasampo/
WarSampo – Finnish World War II on the Semantic Web\textsuperscript{23} (online since 2015 with several new perspectives published in 2016–2019) \cite{16} is a popular Finnish service that has had 740 000 users. It introduced Principle (6) 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. WarSampo application got in 2017 the LODLAM Open Data Prize in Venice.

Interest in WarSampo lead to a new Sampo the same application domain of war history: WarVictimSampo (1914–1922)\textsuperscript{24} (online since 2019) \cite{40} 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, to consider. Based on this activity, a new citizen science project for collecting and maintaining such RDF data is currently underway\textsuperscript{25}.

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

BiographySampo – Finnish Biographies on the Semantic Web\textsuperscript{26} (online since 2018) \cite{19} is yet another popular service with tens of thousands of users. It harnessed Principles (3) and (5) into the Sampo model, with a focus on supporting biographical and prosopographical research and data analyses. 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 some 16 external data sources and by reasoning, e.g., family relations \cite{30} and serendipitous connections between people and places \cite{20}.

The idea of publishing textual biographies as structured LOD for data exploration and analysis was also developed in the Sampos Norssit Alumni \cite{18} and U.S. Congress Prosopographer \cite{37}. AcademySampo\textsuperscript{27} (online since 2021) \cite{30} 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\textsuperscript{28} (online since 2019) \cite{25} 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

\textsuperscript{23} https://seco.cs.aalto.fi/projects/sotasampo/en/
\textsuperscript{24} https://seco.cs.aalto.fi/projects/sotasurmat/
\textsuperscript{25} https://seco.cs.aalto.fi/projects/sotasampo/citizens/en/
\textsuperscript{26} https://seco.cs.aalto.fi/projects/biografiasampo/en/
\textsuperscript{27} https://seco.cs.aalto.fi/projects/akatemiasampo/en/
\textsuperscript{28} https://seco.cs.aalto.fi/projects/nimisampo/en/
register (ca. 800,000 places) of the National Survey of Finland as Linked Open Data. Furthermore, the Thesaurus of Geographical Names (TGN) of Getty Research via its SPARL endpoint is re-used, as well as various map services, including a collection historical maps of Finland published as part of WarSampo. The NameSampo project developed the first version of the Sampo-UI framework [24] that has been used after this is all Sampo, supporting implementation of Principles 3–5 from an UI point of view. Sampo-UI has also been re-used in Norway by the Norwegian Language Collections for creating a national service similar to NameSampo: Norske stedsnavn. The Sampo-UI framework, available openly in Github, has also been re-used in a commercial setting.

**Mapping Manuscript Migrations (MMM)** (online since 2020) [17,27] is a Sampo, in spite of it name, based on metadata about some 220,000 pre-modern manuscripts from the Schoenberg Database of Manuscripts in the U.S., Medieval Manuscripts in Oxford Libraries in the U.K., and the Bibale database in France. MMM is a result of the Trans-Atlantic Digging into Data research programme.

In addition, new Sampo are underway: **FindSampo** [21] is a system and data service for supporting archaeology especially form a citizen science and metal detectorists’ perspectives. **LawSampo** [22] publishes Finnish legislation and case law based on data from the Ministry of Justice in Finland. Both FindSampo and LawSampo portal demonstrators and LOD services already exist and will be published openly online later in 2021. **ParliamentSampo** will publish LOD extracted from the materials of the Parliament of Finland (1907–2021). First version of the LOD service has been created based on the Parliamentary debates and prosopographical data about the politicians.

**LetterSampo** [47] is based on early modern epistolary metadata aggregated in the Early Modern Letters Online (EMLO) service at the Oxford University, the CKCC corpus underlying ePistolarium of the Huygens Institute in the Netherlands, and correspSearch service of the Berlin-Brandenburg Academy of Sciences.

---

29 http://www.getty.edu/research/tools/vocabularies/tgn/
30 https://toponymi.spraksamlingane.no/nb/app
32 See https://sdbm.library.upenn.edu
33 See https://medieval.bodleian.ox.ac.uk
34 The Bibale web service from the Institute for Research and History of Texts (IRHT) in Paris is described in http://bibale.irht.cnrs.fr.
35 https://diggingintodata.org/
36 https://seco.cs.aalto.fi/projects/sualt/
37 https://seco.cs.aalto.fi/projects/lawlod/
40 https://seco.cs.aalto.fi/projects/rrl/
41 http://emlo.bodleian.ox.ac.uk
42 http://ckcc.huygens.knaw.nl/epistolarium/
43 https://correspsearch.net
4 Discussion

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 and has been developed also in various other settings, e.g., in ResearchSpace.

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 in terms of open ontology services and LOD services.

3. The Principle of supporting data analysis and knowledge discovery (P3) based on Big Data is fundamental in, e.g., distant reading, Humanities Computing, and Digital Humanities 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. 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 for epistolary data. The idea of using multiple perspectives has also been studied as an approach in decision making.

5. Regarding Principle P5, faceted search, also known as “view-based search” and “dynamic ontologies”, is a well-known paradigm for explorative search and browsing in computer science and information retrieval, based on S. R. Ranganathan’s original ideas of faceted classification in Library Science. The two step usage model in Sampo model is also used as a general research method in prosopographical research.

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. 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 1–6 and in operationalizing them as 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 a few months after the seminal Semantic Web paper was published in Scientific American 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) [15].

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 [28], 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 [29] 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 [38], 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 finding and explaining serendipitous semantic connections between places and persons to the end user [19,20].

Acknowledgments Tens of people48 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 part of the Semantic Parliament (ParliamentSampo) project of the Academy of Finland, the EU project InTaVia: In/Tangible European Heritage49, and is related to the EU COST action Nexus Linguarum50 on linguistic data science.

References


48 https://seco.cs.aalto.fi/people/
49 https://intavia.eu/
50 https://nexuslinguarum.eu/the-action