How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web

Keynote at the DCMI 2021 Conference

Eero Hyvönen

Aalto University, Department of Computer Science and
University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)
Semantic Computing Research Group (SeCo)
https://seco.cs.aalto.fi/u/eahyvone/

Abstract. The vision of the Semantic Wed is to build the global Web of Data (Giant Global Graph, GGG) for machines to use: based on this an interoperable and intelligent transnational WWW for humans can be created cost-efficiently. This paper addresses this vision on a cross-domain national level, as in practice the data available are often related to each other within national contexts and application domains, and are represented using national languages, metadata models, vocabularies, and local conventions. The paper presents lessons learned in Finland on developing and deploying a cross-domain national ontology and Linked (Open) Data (LOD) infrastructure. To test and demonstrate the infrastructure, a series of 16 semantic portals and LOD services in use have been created using a model that has evolved gradually in 2002–2021. They cover a wide range of application domains and have attracted millions of users in total suggesting feasibility of the proposed model. This work shows a shift of focus in research on semantic portals from data aggregation and exploration systems to systems supporting research with data analytic tools, and finally to automatic knowledge discovery and Artificial Intelligence.

Keywords: Semantic Web, Linked Data, Infrastructures, Portals

1 Extending the Layer Cake Model

The Semantic Web (SW) sees the WWW as an interlinked collection of data (Web of Data) instead of only a space of interlinked hypertext documents, Web of Pages. The idea was proposed in the 90's by Tim Berners-Lee [2] and first recommendations (standards) for the SW[1], such as the Resource Description Framework (RDF), were developed before the millenium. The SW recommendations constitute the W3C "layer cake model" [11,7] on top of XML, the lingua franca of the WWW, and lay out a new basis of shared semantics for interoperability of data. Founded on using first order predicate logic, the semantics of the SW [17] are independent of application domains

[1] https://www.w3.org/standards/semanticweb/

and natural languages. This makes the model suitable for dealing with the versatile data underlying the Web.

However, the layer cake model is not enough: domain and application specific infrastructures based on shared W3C standards and best practices are needed, too. These can focus on specific domains, such as medicine, biology, cultural heritage, or geography on an international level. However, in practice one also has to deal with national level issues and data available that are represented using national languages, data models, vocabularies, and are created using conventions of local legacy systems. For example, Cultural Heritage (CH) data in different countries is often nationally specific calling for adapted local solutions for representing and using the data.

This paper concerns the question: *How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web*. This problem is addressed by presenting, discussing, and evaluating approaches and living laboratory experiments developed in Finland during the last twenty years 2001–2021. Presenting lessons learned in this particular endeavour is hopefully useful in a more general setting, as similar challenges are likely to be faced in other countries, too.

In Section 2, elements needed for a national SW infrastructure are first introduced. The idea and lessons learned in developing a national ontology and a LOD infrastructure are then presented in Section 3 and Section 4, respectively. After this, applications of the infrastructures are discussed as a proof-of-concept: a model is presented that has been used for creating a series of 16 in-use portals and data services on the SW. Finally, contributions of the work are summarized and related works discussed. This paper presents the first consolidated account of this line of research and development, summarizing works reported before in some 450 papers[2].

2 Elements of SW Infrastructure

Fig. 1 depicts components that are needed in the developing a national SW infrastructure—according to the experiences to be reported in this paper. The system is based on domain agnostic W3C *Web Standards and Best Practices* (on the left below in the figure) of publishing Linked Data[3] [14]. *Data Models* are needed for representing metadata and knowledge of different applications domains, populated by resources taken from shared domain *Ontologies and Ontology Services* for interoperability. The ontologies should be made openly available and easy to access for interoperability and re-use, based on shared ontology services/libraries; cf. [4] for a survey of such systems. In the same vein, data services for publishing LD datasets, preferably using, e.g., open Creative Commons licenses, are needed for making re-use of data possible and easy. Also *Applications* of Linked Data are part of the infrastructure connecting the system to its end users. For making all this possible, *Software Tools* are needed for aggregating the distributed heterogeneous data from legacy and other data silos involved, and for extracting and linking (disambiguating) entities and relations from data records and textual descriptions [54]. Also tools for data publishing and analysis are needed, as well as tooling for developing new applications for the end users.

[2] http://seco.cs.aalto.fi/publications/

[3] https://www.w3.org/TR/ld-bp/

For developing, maintaining, and using the infrastructure in a sustainable way a *Human Infrastructure* is needed (on the left in Fig. 1). This involves, e.g., educating people about the technology, and production of documentations and learning materials for the community using national languages. In the Finnish case, for example, online materials have been created, the first Finnish text book about SW was produced, and hackathons[4] were organized on using the data and tools. SW courses were also included in university curricula.

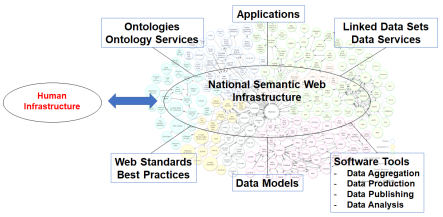

Fig. 1. Elements needed for a national SW intrastructure

3 Ontology Infrastructure

In early 2000's, the focus in SW research was on ontologies [64], arguably the "silver bullet" of the SW [8]. Accordingly, a series of projects called "FinnONTO" in 2003–2012 were started[5] in Finland.

A National Effort The goal in the FinnONTO initiative [20] was to develop a national ontology and content infrastructure, based on W3C standards, that would be cross-domain, multilingual (Finnish, Swedish, and English) and openly available. The consortium behind the initiative included finally some 50 companies and public organizations that represented a wide spectrum of functions of the society, including libraries, health organizations, cultural institutions, government, industry, media, and education.

The initiative produced 1) metadata models contributing to national standards[6], 2) ontologies [70] to be used for populating the metadata models, 3) a living laboratory called ONKI of public ontology services [74], and 4) tools for metadata creation and application development, such as Skosify [65] for SKOS vocabulary quality assessment and SAHA editor [43] for managing RDF repositories. The infrastructure was tested by using it in case studies in different application domains, including e-culture [23,51], e-health [66], e-government [62], e-learning [44], and e-commerce [46].

[4] E.g., as part of the Helsinki DH hackathon series https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/helsinki-digital-humanities-hackathon

[5] https://seco.cs.aalto.fi/projects/finnonto/

[6] Such as the Public Recommendation for Geographic Metadata, Ministry of Internal Affairs, http://www.jhs-suositukset.fi/suomi/jhs158.

A central goal FinnONTO was to create an interlinked cloud of national ontologies [9] based on existing thesauri that were already used in different areas of the society. The rationale for this was that metadata available in national databases had already been catalogues using these thesauri, which would make it much easier to develop applications. According to the FinnONTO vision, the ontologies should be served not only through human readable browser interfaces, but also as centrally managed national ontology services using REST APIs. In this way, common functionalities of the services, such as (semantic) autocompletion, URI fetching, and query expansion, could be shared on a national level, and everybody would get access to up-to-date versions of the ontologies. This kind of collaboration would be cost-efficient on a national level and gradually lead to better interoperability of the data catalogued in different organizations. Availability of centralized services is needed especially for smaller organizations that do not have much expertise and resources for developing their own web services.

From Thesauri to Ontologies The FinnONTO project transformed the key national thesauri used in Finland into light-weight ontologies listed in Table 1. The transformation process was more ambitious than just transforming the traditional standard thesaurus format [1] into an RDF-based model. The thesauri were developed semantically a bit forward, using the OntoClean methodology [12] and RDFS, in the following ways [20]: 1) Multiple meanings of thesauri terms were disambiguated and relocated in `rdfs:subClassOf` hierarchies. For example, the concept of *child* could refer to the class of young people, to a family relation, or a social class, that should be in different places of the ontology. 2) The thesauri involved did not differentiate whether the standard Broader Term (BT) relation [1] means part-of or hypernymy relation. This distinction was crafted manually in the ontologies. 3) The `rdfs:subClassOf` hierarchies were completed: all concepts were given at least one superclass (except the roots). 3) Inheritance of instanceship over subclass hierarchies was checked as specified in RDFS, so the hierarchies could be used for reasoning in, e.g., query expansion.

Linked Ontology Cloud KOKO The ontologies in Fig. 1 share lots of similar concepts. For example, in Table 2 based on [9], shared concepts between five ontologies of Table 1 are listed. The by far largest and most used thesaurus and ontology YSO (27 200 concepts) of the National Library shared lots of concepts with all other ontologies, often more than 50%. This suggested that the ontologies should be linked together using YSO as the top ontology. This resulted in creating the Finnish linked ontology cloud called KOKO[7].

Lessons Learned An initial problem to be solved in FinnONTO was that large cross-domain thesauri, especially the General Finnish Thesaurus, could not anymore be maintained easily by its management team. Even if the team included people from different fields, the terminology related to specific areas needed deeper domain specific expertise than was available. Developing the interlinked KOKO ontology cloud mitigates the problem by distributing work on specific concepts to collaborative, domain specific ontology developer teams. However, in this model new problems arise pertaining to maintaining the linked ontology cloud and to coordinating the collaboration network [9]. These new challenges are now being tackled by the Finto collaboration network coordinated by the National Library. FinnONTO pointed out that lots of redundant work

[7] https://finto.fi/koko/fi/

Ontology	Domain	# of concepts
YSO	General upper ontology (GUO)	27 200
AFO	Agriculture and forestry	7000
JUHO	Government	6300
KAUNO	Literature	5000
KITO	Literary research	850
KTO	Linguistics	900
KULO	Cultural research	1500
LIITO	Economics	3000
MAO	Museum artifacts	6800
MERO	Seafaring	1300
MUSO	Music	1000
PUHO	Military	2000
TAO	Design	3000
TERO	Health	6500
TSR	Working and employment	5100
VALO	Photography	2000

Table 1. The linked ontologies of the KOKO cloud

Ontology	AFO	JUHO	KAUNO	MERO	TERO
AFO	100%	8%	2%	3%	25%
JUHO	7%	100%	16%	5%	40%
KAUNO	2%	12%	100%	1%	28%
MERO	0%	1%	0%	100%	2%
TERO	23%	41%	36%	13%	100%

Table 2. Shared concepts in five KOKO ontologies

had been done in developing the thesauri in Finland as they shared lots similar concepts with each other. In the new, more coordinated KOKO model, redundant work can be better eliminated.

A challenge encountered in the ontologization process was that organizing the concepts into class hierarchies cannot in many cases represent correctly the meaning of the original terms that can be complex and fuzzy. The world cannot be represented fully using ontologies and there can be several ways in which this can be done. In spite of such challenges, the idea of adding a little semantics seems to be a better option that continuing using the original thesauri. A strategic choice made in FinnONTO was to follow the wisdom articulated by Jim Hendler already in the late 90's in the SHOE project[8]: *A little semantics goes a long way.* In our case, the thesauri semantics were refined only a little using RDFS for interoperability and to help development of web services. However, already this was a hand-full of work, as thousands of terms in the thesauri had to be manually checked and refined [61].

A mundane challenge of developing large vocabularies, at least in Finland, is how to convince the funding organizations, year after year, that this never ending work should be supported in a sustainable manner, not only as separate short-time projects. In our case, it took some ten years of project-based work before the KOKO ontology infrastructure and Finto services could be funded in a more sustainable way by two Finnish ministries. The strategy taken in FinnONTO was to move forward in baby steps, and after each step show a demonstrator on how the ontologies can be applied in practise for creating something useful, such as the semantic Sampo portals to be discussed later on in this paper.

The idea of creating a "living laboratory" of ONKI ontology services [74,70] on the Web turned out to be useful for deploying the infrastructure. The participating FinnONTO organizations were supported by the project in connecting their legacy systems to the APIs of ONKI for testing and evaluating the services. Finally, the "point of no return" was reached where pulling off the plug of the services was not an option anymore as the number of ONKI API users were counted already in hundreds.

The FinnONTO projects 2003–2012 started with a smallish one-year project, but eventually grew into a national effort of substantial size on the Finnish scale with tens of funding organizations involved. A reason for this was that in addition to public organizations, such as museums, libraries, and archives, also companies got interested in the technology, which convinced the main funding organization Tekes (called today Business Finland) that something useful and of monetary value is happening related to semantic web technologies. It is usually easier to get funding for technology development than for research in humanities.

The KOKO ontologies are based on keyword thesauri whose terms usually correspond to the classes. FinnONTO worked also on various "instance-based" ontologies, such as national geogazetteers, person and organization registries, biological taxonomies of species [56,71], and nomenclatures and terminologies of medicine, such as MESH[9]. Creating a national ontology infrastructure is a never-ending job and goes

[8] https://www.cs.rpi.edu/ hendler/LittleSemanticsWeb.html
[9] https://finto.fi/mesh/fi/

on, e.g., in the Linked Open Infrastructure for Digital Humanities initiative in Finland initiative[10] [24].

When developing ontology-based applications in FinnONTO, much of the time of the developers was "waisted" in cleaning and aligning the data from different organizations for interoperability. Obviously, it would be more cost efficient do this work already when cataloging the data using ontology services. This would also enhance the quality of the linked data, which is a critical problem [75] on the SW. The local cataloguers know best their own data and should have the best interest data quality. The motto for the FinnONTO work was therefore taken from a wisdom of Albert Einstein: *Intellectuals solve problems – geniuses prevent them.*; a key goal of FinnONTO was to prevent interoperability problems rather than to solve them afterwards when the damage was already done in cataloguing [21].

A major outcome of FinnONTO was the ONKI ontology server with its ontologies [74] that were publihied first in 2008. A central part of ONKI, ONKI Light service[11], was developed later and deployed in 2014 [67] by the National Library of Finland as the national Finto.fi service[12]. ONKI Light finally evolved in the open source Skosmos tool[13] in use in several other organizations, too. ONKI Light was based on a SPARQL endpoint. This idea was to separate the data service fully from the user interface. This idea turned later useful when developing the Sampo model and Sampo-UI tool for semantic portals to be discussed later in this article.

Finto has grown into a popular national free service. In 2019 it was used by 280 000 different users and its APIs were called 32 million times. The users include, e.g., museums, whose cataloging system get their keywords with URI identifiers from Finto. These developments suggest that the fundamental ideas of FinnONTO are feasible; they have actually made a paradigm change in Finland in developing and using linked light-weight ontologies on a national level instead of thesauri.

4 From 5-star to 7-star model

A SW infrastructure (cf. Fig. 1) should include a platform for publishing datasets and (re)using them via web services. A key component in LD publishing is the SPAQL endpoint, but the platform should also support other functions [14]. The Linked Data Finland service LDF.fi[14] [36] was therefore developed in follow-up projects of FinnONTO.

LDF.fi has two user-groups: 1) For application developers, LDF.fi provides SPARQL endpoints and a suite of standard Linked Data (LD) services, including content negotiation, APIs for downloading datasets, LD browsing and editing, and additional tools for, e.g., data documentation and visualization. 2) For data publishers, the idea is to support and automate the data publishing process in the following way: The publisher creates a service description of the dataset and its schemas, using an extended version of

[10] https://seco.cs.aalto.fi/projects/lodi4dh/
[11] https://seco.cs.aalto.fi/services/onkilight/
[12] Available at: https.//finto.fi
[13] https://skosmos.org/
[14] https://ldf.fi

the W3C Service Descripton recommendation[15]. Based on such metadata, LDF.fi then 1) automatically sets up the technical services, 2) generates a dataset "homepage" that explains the dataset, schemas, and 3) provides additional related services for querying, documenting, inspecting, and validating the data.

Linked data publications on the SW are typically evaluated with the W3C "5-star model"[16], using a quality scale analogous to evaluating hotels. In LDF.fi, the 5-star model is extended to a 7-star model: there are nowadays also a few 7 star hotels around[17]. The 6th star is given to a data publication if it includes not only the 5-star data but also the schemas of the data with documentation. This makes re-use of data easier. The 7th star is give to a data publication, if the publication includes some kind of evaluation that the data actually conforms to the provided schemas using, e.g., the SHACL Shapes Constraint Language[18] or ShEx Shape Expressions[19] [45]. The idea here is to encourage publishers to publish high quality data as data quality of LD is a severe issue on the SW.

Schemas can be documented automatically in LDF.fi for the human reader using a schema documentation generator, in our case SpecGen[20]. Datasets in the LD world often use schemas (vocabularies) for which definitions or descriptions are not available, but are embedded in the data itself. In order to find out how schemas are actually used in a dataset, including both published and unpublished schemas, a service vocab.at[21] was created that analyzes a given dataset from this perspective and creates an HTML document listing, e.g., statistics of vobaculary usage and raising up issues detected, e.g., if an IRI is not dereferenceable. The input for vocab.at is either an RDF file, a SPARQL endpoint, or an HTML page with embedded RDFa markup.

LDF.fi is implemented by a combination of the Fuseki SPARQL server[22] for storing the primary data and a Varnish Cache web application accelerator[23] for routing URIs, content negotiation, and caching. For simple deployment of applications with a data service (cf., e.g., the MMM system [28]) a microservice architecture with Docker containers[24] is used. Each individual component (the application, Varnish, and Fuseki) is run in its own dedicated container, making the deployment of the services easy due to installation of software dependencies in isolated environments. This enhances the portability of the services. The server environment of LDF.fi is provided by the CSC – IT Center for Science, a company of the Ministry of Education and Culture of Finland providing computational infrastructures for the national universities.

Lessons learned The Linked Data Finland platform has turned out to be useful for data-analytic research purposes and in developing applications (cf. Section 5). LDf.fi

[15] http://www.w3.org/TR/sparql11-service-description/
[16] https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/
[17] Such as the Burj Al Arab in United Arab Emirates
[18] https://www.w3.org/TR/shacl/
[19] https://shex.io/
[20] https://bitbucket.org/wikier/specgen/wiki/Home
[21] http://vocab.at
[22] https://jena.apache.org/documentation/fuseki2/
[23] https://varnish-cache.org
[24] https://www.docker.com

has been used for publishing some 100 linked datasets. Some of them are in use in the Sampo portals to be discussed in the next section and via SPARQL querying combined with query editing and scripting tools using the open CC BY 4.0 license. Some datasets are used only internally in related research projects, and for some datasets licensing policy of the data owners prohibits open use. LDF.fi hosts several instance-based ontologies, too, such as an RDF-version of the ca. 800 000 official Finnish geographical places based on data of the National Survey.

The LDF.fi service is still maintained by Aalto University and University of Helsinki that developed it on an academic project basis, but with the hope that some day it will be deployed and be maintained in a more sustainable way—this is at least what happened to the related ONKI/Finto ontology services. A step towards this is that in 2020 the concept of providing Linked Open Data services on a national level and LDF.fi were accepted on the new research infrastructure roadmap of the Academy of Finland as part of the larger initiative FIN-CLARIAH[25]. Here the idea is to combine—on a national level as in the CLARIAH initiative in the Netherlands—the work related to the pan-European infrastructures CLARIN[26], the research infrastructure for language as social and cultural data, and DARIAH[27], the infrastructure for arts & humanities scholars.

5 Applying the SW Infrastructure

When developing the Finnish SW infrastructure, applications testing and demonstrating its usability were constantly developed. This work evolved gradually into a general model for developing semantic portals, called the *Sampo Model*, and the *Sampo Series* of semantic portals and data services[28] [26]. The novelty of the Sampo model[29] lays in its attempt to formulate a set of re-usable design principles or guidelines for creating semantic portals, especially for Cultural Heritage applications and Digital Humanities research [10]. Based on six principles, the 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 [16].

5.1 Sampo Model Principles

The six Sampo model principles P1–P6 are described and motivated in more detail below.

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 [22]. The

[25] https://www.kielipankki.fi/organization/fin-clariah/

[26] https://clarin.eu

[27] https://dariah.eu

[28] See https://seco.cs.aalto.fi/applications/sampo/ for a complete list of "Sampo portals", videos, and further information.

[29] 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.

local data are harmonized and enriched with each other by linking and reasoning. In this model everybody can arguably 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.

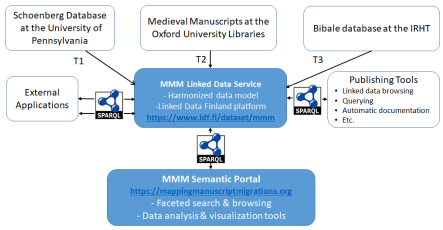

Fig. 2. Publishing and using heterogeneous distributed data in the MMM Sampo system

Fig. 2 depicts as an example of how the principle P1 was used in the Mapping Manuscript Migrations (MMM) system [29]. 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 (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 [39] 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).

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. Most Sampo systems make use of the FinnONTO ontology infrastructure and the LDF.fi LOD services.

P3. Support data analysis and knowledge discovery in addition to data exploration Three generations of semantic portals for Digital Humanities can be identified [25]. 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 AI: 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. The Sampo model aims not only at data publishing with search and data exploration [53] but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analysing, and even solving research problems in interactive ways.

P4. Provide multiple perspectives 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 [23] when dealing with Big Data. The application perspectives can be provided on the landing page of the Sampo portal system, and they enrich each other by data linking. Also 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.

Fig. 3. MMM Sampo with five application perspectives

For example, Fig. 3 depicts the landing page of the MMM Sampo [27] with five interlinked application perspectives on finding and analyzing Manuscripts, Works, Events, Actors, and Places in the underlying LOD service data.

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 [19,69,73]. Second, the target group is visualized or analyzed by using ready-to-use data analytic 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 [37].

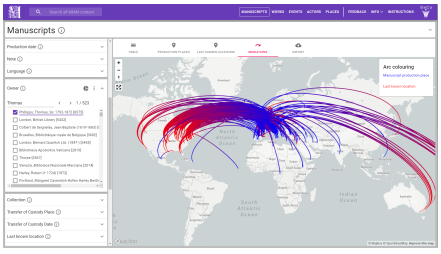

Fig. 4. Migrations of manuscripts owned by Sir Thomas Phillipps (1792–1872) from the place of production (blue end of an arc) to the last known location (red end of the arc)

This idea is illustrated in the MMM portal screenshot of Fig. 4 for the 8575 manuscripts owned by the collector Sir Thomas Phillipps (1792–1872). The manuscripts were filter out by a selection in the collection owner facet on the left. By then selecting the visualization tab Migrations on the right it can be seen on a map how the manuscripts have migrated around the world since medieval times. This visualization is an answer to one of the original research question in manuscript studies set when starting the MMM project [3].

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[30] [59] interface for SPARQL querying and visualizing the results can be used, or Python scripting in Google Colab[31] and Jupyter notebooks[32] [68].

The Sampo model principles above are compatible with the FAIR principles for creating Findable, Accessible, Interoperable, and Re-usable data[33], but were developed in the context of publishing and using Cultural Heritage LOD. The model can, however, be applied in other domains, too. An example of this is the HealthFinland system [66] for health promotion information, that was deployed by the National Institute for Health and Welfare in Finland[34].

Sampo Framework and Application Layers The principles P1–P6 can be used directly for creating semantic portals. However, its is also possible to apply them first to create an application domain specific framework and reuse it for developing different related application instances, which is arguably cost-efficient. Fig. 5 illustrates the idea with LetterSampo and FindSampo as an example. The highest conceptual layer includes the Sampo model with its principles based on domain agnostic, logical SW standards of the W3C and Linked Data publishing principles. On the next domain specific level, model level solutions and principles are applied to create a domain specific framework by using a domain specific data model that can be populated using domain specific vocabularies and ontologies (e.g., persons sending /receiving letters, archaeological object types, historical places, etc.). This layer includes also a domain specific template designed using the Sampo-UI framework [37] that can be copied and used as a starting point for creating application instances. The template tells, e.g., what thematic application perspectives, data-analysis tools, and ready-to-use UI components are available in this application domain. The figure depicts the LetterSampo framework for epistolary domain [32], applied to three international datasets of letters, and the FindSampo framework [34] for archaeological finds, applied to a dataset of the National Museum of Finland and a dataset of the British Museum. Finally, applications can be created by adding in specific datasets into the framework, by creating a Sampo-UI implementa-

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

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

[32] https://jupyter.org

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

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

tion of the portal interface, and by publishing the data in a Linked Data service with a SPARQL endpoint.

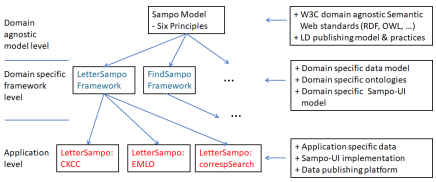

Fig. 5. Three conceptual layers for creating Sampo portals for re-using generic upper level solutions in applications.

6 Applications and Services in Use

The Sampo model has evolved gradually in 2002–2021 via lessons learned in developing a series of semantic portals and LOD services. This section overviews shortly a selection of these systems listed in Table 3, with a focus on Digital Humanities. 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, these systems are described shortly in order to provide a proof-of-concept of the Sampo model and the usability of the underlying national infrastructure.

MuseumFinland – Finnish Museums on the Semantic Web[35] (online since 2004) [18] was the first Sampo. It introduced principle P1 by aggregating and publishing heterogeneous, distributed artefact collection data from Finnish museums.[36]

CultureSampo – Finnish Culture on the Semantic Web 2.0[37] (online since 2009) [23,51] 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[38] (online since 2011) [49] 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 libraries, and is used by ca. 2 million users in a year.

[35] https://museosuomi.fi
[36] This application got the Semantic Web Challenge Award at the ISWC 2004 conference.
[37] https://seco.cs.aalto.fi/applications/kulttuurisampo/
[38] https://seco.cs.aalto.fi/applications/kirjasampo/

WarSampo – Finnish World War II on the Semantic Web[39] (online since 2015 with several new perspectives published in 2016–2019) [27] is a popular Finnish service that has had 857 000 users. It introduced principle P6 into the Sampo model. The portal and its data service provides information about over 100 000 casualties and significant soldiers of the Second World War in Finland. The dataset includes various graphs, such as 160 000 authentic photographs from the fronts, 26 000 war diaries, historical maps, memoir articles of soldiers, etc., constituting small a LOD cloud of its own and an infrastructure for Finnish WW2 data [41].[40] Interest in WarSampo lead to a new Sampo in the same application domain of war history: **WarVictimSampo (1914–1922)**[41] (online since 2019) [58] publishes data about the deaths and battles of the Finnish Civil War 1918 and related wars.

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 applications discussed below.

BiographySampo – Biographies on the Semantic Web[42] (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 [47] 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** [30] and **U.S. Congress Prosopographer** [55]. **AcademySampo**[43] (online since 2021) [47] 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[44] (online since 2019) [38] 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)[45] 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.

[39] https://seco.cs.aalto.fi/projects/sotasampo/en/
[40] WarSampo application got in 2017 the LODLAM Open Data Prize in Venice.
[41] https://seco.cs.aalto.fi/projects/sotasurmat/
[42] https://seco.cs.aalto.fi/projects/biografiasampo/en/
[43] https://seco.cs.aalto.fi/projects/akatemiasampo/en/
[44] https://seco.cs.aalto.fi/projects/nimisampo/en/
[45] http://www.getty.edu/research/tools/vocabularies/tgn/

The NameSampo project developed, based on the SPARQL Faceter tool [40] used in many earlier Sampos, the first version of the Sampo-UI framework [37] that has been used after this is in 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[46]. The Sampo-UI framework, available in Github[47], has also been re-used in a commercial setting.

Mapping Manuscript Migrations (MMM)[48] (online since 2020) [28,39] is a Sampo based on metadata about some 220 000 pre-modern manuscripts from the Schoenberg Database of Manuscripts[49] in the U.S., Medieval Manuscripts in Oxford University Libraries[50] in the U.K., and the Bibale[51] database of IRHT in France.

FindSampo[52] [34] (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**[53] [35] publishes Finnish legislation and case law based on data from the Ministry of Justice in Finland. **ParliamentSampo**[54] publishes LOD extracted from the materials of the Parliament of Finland (1907–2021), including knowledge graphs about over 900 000 Parliamentary debate speeches [63] and prosopographical data about the politicians' networks [48] in 1907–2021. **LetterSampo**[55] [72] is based on early modern epistolary metadata aggregated in the Early Modern Letters Online (EMLO) service[56] at the Oxford University, the CKCC corpus underlying ePistolarium[57] of the Huygens Institute in the Netherlands, and correspSearch[58] service of the Berlin-Brandenburg Academy of Sciences.

7 Contributions and Related Works

This paper addressed challenges of extending the SW layer cake model for creating ontology and LOD infrastructures especially on national and domain specific levels. Lessons learned in developing Finnish ontology and linked data services 2001–2021 were discussed. This work has utilized methods of design science [52,15,57] and action research [6], where the idea is to design artifacts, evaluate their value and utility, and to provide improvements in solutions. Rather than creating theoretical knowledge,

[46] https://toponymi.spraksamlingane.no/nb/app
[47] https://github.com/SemanticComputing/sampo-ui
[48] https://seco.cs.aalto.fi/projects/mmm/
[49] See https://sdbm.library.upenn.edu
[50] See https://medieval.bodleian.ox.ac.uk
[51] http://bibale.irht.cnrs.fr
[52] https://seco.cs.aalto.fi/projects/sualt/
[53] https://seco.cs.aalto.fi/projects/lawlod/
[54] https://seco.cs.aalto.fi/projects/semparl/en/
[55] https://seco.cs.aalto.fi/projects/rrl/
[56] http://emlo.bodleian.ox.ac.uk
[57] http://ckcc.huygens.knaw.nl/epistolarium/
[58] https://correspsearch.net

design science applies knowledge. In this paper, infrastructure elements were designed, implemented, and applied to create the Sampo series of data services and portals as a proof-of-concept. They have had up to millions of end users (Table 3), which suggests feasibility of the national infrastructures presented. The line of R&D presented is unique in its focus on different domains on a national level, longevity, the Sampo model, and series of evolving applications in-use on the Semantic Web based on it.

The idea of ONKI/Finto ontology services was inspired by early ideas of ontology libraries. In contrast to current related ontology library systems [5] that typically focus on particular application domains, ONKI and Finto aimed at being a cross-domain ontology service on a national level. For example, the BioPortal [60] of Stanford University is focused on publishing biomedical ontologies.

There are lots of LOD services and SPARQL endpoints around[59]. The novelty of the LDF.fi service lays on its 7-star model and the idea of integrating the data service with various online tools as well as leaning materials to support data reuse. Instead of being a focused data service for particular data, such as DBpedia for Wikipedias, the LDF.fi platform aims at being a cross-domain platform of datasets on a national level. The main application area of the presented infrastructure has been Cultural Heritage and Digital Humanities [10], especially in the 10's, although also systems for, e.g., e-health, e-government, and e-learning were developed.

During the past 20 years, the SW has evolved in phases [16] with a focus first on ontologies [64], then on Linked (Open) Data [14], and today on Knowledge Graphs (KG) [13]. The Sampo series reflects this development by showing a shift of focus form data publishing, based on shared ontologies and metadata vocabularies[60] (1. generation portals), to supporting the end-users of KGs with seamlessly integrated data-analytic tools and visualizations needed in areas such as Digital Humanities (2. generation systems). However, the series has also taken first steps forward towards 3. generation portals that can solve problems for the end users based on knowledge discovery, Artificial Intelligence, and computational creativity [25]. There are lots of related works pertaining to the different Sampo systems overviewed in this paper. Discussing them is beyond the scope of this paper, but pointers to such works can be found in the referenced research papers about the Sampos.

The experiences reported in this paper indicate that creating and using a national semantic web infrastructure is useful from the data producers' and data users' points of view. However, creating and using linked data has its own challenges, too. More collaboration and agreements on data models and ontologies are needed for interoperability between the data producers, which complicates the publication process. Integration of SW technologies with legacy systems may be challenging, and there is lack of IT personnel competent in using SW technologies and tools. Creating linked data manually is costly but automatic methods may not be available and automation lowers data quality. Using structured semantic data and making the knowledge structures explicit to the end user in the UI calls for a new kind of digital data literacy and source criticism[61] from the end user [42,50]. What the underlying data actually means is not always clear

[59] https://www.w3.org/wiki/SparqlEndpoints
[60] https://lov.linkeddata.es/dataset/lov/
[61] https://ranke2.uni.lu/define-dsc/#%20,%20Universit%C3%A9%20du%20Luxembourg

and issues of Big Data quality, such as completeness, veracity, skewness, uncertainty, fuzziness, and errors of data arise.

References

1. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus Construction and Use: A Practical Manual. Aslib IMI (2000)
2. Berners-Lee, T., Fischetti, M., Dertouzos, M.L.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. Harper San Francisco, 1st edn. (1999)
3. Burrows, T., Emery, D., Fraas, M., Hyvönen, E., Ikkala, E., Koho, M., Lewis, D., Morrison, A., Page, K., Ransom, L., Thomson, E., Tuominen, J., Velios, A., Wijsman, H.: Mapping Manuscript Migrations: Digging into data for researching the history and provenance of medieval and renaissance manuscripts (white paper) (August 2020), https://diggingintodata.org/file/1281/download?token=x59u8fFQ
4. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web **11**, 96–111 (2012)
5. d'Aquin, M., Noy, N.F.: Where to publish and find ontologies? A survey of ontology libraries. Journal of Web Semantics First Look **11**(0) (2012). https://doi.org/10.2139/ssrn.3198941
6. Davison, R.M., Martinsons, M.G., Kock, N.: Principles of canonical action research. Information Systems Journal **14**(1), 65–86 (2004). https://doi.org/10.1111/j.1365-2575.2004.00162.x
7. Domingue, J., Fensel, D., Hendler, J.A.: Introduction to the semantic web technologies. In: Domingue, J., Fensel, D., Hendler, J.A. (eds.) Handbook of Semantic Web Technologies, pp. 1–41. Springer (2011). https://doi.org/10.1007/978-3-540-92913-0_1
8. Fensel, D.: Ontologies: Silver bullet for knowledge management and electronic commerce (2nd Edition). Springer (2004)
9. Frosterus, M., Tuominen, J., Pessala, S., Hyvönen, E.: Linked open ontology cloud: managing a system of interlinked cross-domain light-weight ontologies. International Journal of Metadata, Semantics and Ontologies **10**(3), 189–201 (2015), http://dx.doi.org/10.1504/IJMSO.2015.073879
10. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015), https://doi.org/10.1017/CBO9781139003865
11. Goebel, R., Zilles, S., Ringlstetter, C., Dengel, A.R., Grimnes, G.A.: What is the role of the semantic layer cake for guiding the use of knowledge representation and machine learning in the development of the semantic web? In: AAAI Spring Symposium: Symbiotic Relationships between Semantic Web and Knowledge Engineering (2008)
12. Guarino, N., Welty, C.: Evaluating ontological decisions with OntoClean. Communications of the ACM **45**(2), 61–65 (2002)
13. Gutierrez, C., Sequeda, J.F.: Knowledge graphs. Communications of the ACM **6**(3), 96–104 (March 2021). https://doi.org/10.1145/3418294
14. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Morgan & Claypool, Palo Alto, California (2011), http://linkeddatabook.com/editions/1.0/
15. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly: Management Information Systems **28**(1), 75–105 (2004). https://doi.org/10.2307/25148625
16. Hitzler, P.: A review of the semantic web field. Commun. ACM **64**(2), 76–83 (Jan 2021). https://doi.org/10.1145/3397512
17. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer (2010)

18. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web. Journal of Web Semantics **3**(2), 224–241 (2005)
19. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology-based techniques to view-based semantic search and browsing. In: Proceedings of the First European Semantic Web Symposium. Springer (2004)
20. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach. In: Proceedings of the ESWC 2008, Tenerife, Spain. pp. 95–109. Springer (2008)
21. Hyvönen, E.: Preventing interoperability problems instead of solving them. Semantic Web – Interoperability, Usability, Applicability **1**(1–2), 33–37 (2010)
22. Hyvönen, E.: Publishing and using cultural heritage linked data on the Semantic Web. Morgan & Claypool, Palo Alto, California (2012)
23. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., Nyberg, K.: 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)
24. Hyvönen, E.: 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. pp. 254–259. CEUR Workshop Proceedings, vol. 2612 (October 2020), http://ceur-ws.org/Vol-2612/short10.pdf
25. Hyvönen, E.: Using the Semantic Web in Digital Humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web – Interoperability, Usability, Applicability **11**(1), 187–193 (2020)
26. Hyvönen, E.: Digital humanities on the semantic web: Sampo model and portal series (April 2021), http://semantic-web-journal.org/content/digital-humanities-semantic-web-sampo-model-and-portal-series, submitted
27. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the Second World War history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer (2016)
28. Hyvönen, E., Ikkala, E., Koho, M., Tuominen, J., Burrows, T., Ransom, L., Wijsman, H.: Mapping manuscript migrations on the semantic web: A semantic portal and linked open data service for premodern manuscript research. In: Semantic Web. Proceedings of the The 20th International Semantic Web Conference (ISWC 2021). Springer (2021), forth-coming
29. Hyvönen, E., Ikkala, E., Tuominen, J., Koho, M., Burrows, T., Ransom, L., Wijsman, H.: 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. pp. 220–229. CEUR Workshop Proceedings, Vol-2364 (2019)
30. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: 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). pp. 113–119. Springer (2017), https://link.springer.com/chapter/10.1007/978-3-319-59888-8_9
31. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo – Publishing and enriching biographies on the Semantic Web for digital humanities research. In: Proceedings of the 16th Extended Semantic Web Conference (ESWC 2019). pp. 574–589. Springer (2019)
32. Hyvönen, E., Leskinen, P., Tuominen, J.: Lettersampo – historical letters on the semantic web: A framework and its application to publishing and using epistolary data of the republic of letters (2021), submitted

33. Hyvönen, E., Rantala, H.: Knowledge-based relational search in cultural heritage linked data. Digital Scholarship in the Humanities (DSH) (March 2021), accepted
34. Hyvönen, E., Rantala, H., Ikkala, E., Koho, M., Tuominen, J., Anafi, B., Thomas, S., Wessman, A., Oksanen, E., Rohiola, V., Kuitunen, J., Ryyppö, M.: Citizen science archaeological finds on the semantic web: The FindSampo framework. Antiquity, A Review of World Archaeology (Jan 2021), accepted
35. Hyvönen, E., Tamper, M., Ikkala, E., Sarsa, S., Oksanen, A., Tuominen, J., Hietanen, A.: 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, pp. 110–114. Springer (2020). https://doi.org/10.1007/978-3-030-62327-2_19
36. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: 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. pp. 226–230. Springer (May 2014). https://doi.org/10.1007/978-3-319-11955-7_24
37. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. Semantic Web – Interoperability, Usability, Applicability (2021), accepted
38. Ikkala, E., Tuominen, J., Raunamaa, J., Aalto, T., Ainiala, T., Uusitalo, H., Hyvönen, E.: Namesampo: A linked open data infrastructure and workbench for toponomastic research. In: Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities. pp. 2:1–2:9. GeoHumanities'18, ACM, New York, NY, USA (November 2018). https://doi.org/10.1145/3282933.3282936
39. Koho, M., Burrows, T., Hyvönen, E., Ikkala, E., Page, K., Ransom, L., Tuominen, J., Emery, D., Fraas, M., Heller, B., Lewis, D., Morrison, A., Porte, G., Thomson, E., Velios, A., Wijsman, H.: Harmonizing and publishing heterogeneous pre-modern manuscript metadata as linked open data (2021), accepted, JASIST Special Issue
40. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter – client-side faceted search based on SPARQL. In: Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings (2016), http://www.ceur-ws.org/Vol-1615, vol. 1615
41. Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., Hyvönen, E.: WarSampo knowledge graph: Finland in the second world war as linked open data. Semantic Web – Interoperability, Usability, Applicability **12**(2), 265–278 (Jan 2021). https://doi.org/10.3233/SW-200392
42. Koltay, T.: Data literacy for researchers and data librarians. Journal of Librarianship and Information Science **49**(1), 3–14 (2015). https://doi.org/10.1177/0961000615616450
43. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010) at ESWC 2010. CEUR Workshop Proceedings, Vol. 596 (2010)
44. Känsälä, T., Hyvönen, E.: A semantic view-based portal utilizing Learning Object Metadata. In: 1st Asian Semantic Web Conference (ASWC2006), Semantic Web Applications and Tools Workshop (2004)
45. Labra Gayo, J.E., Prud'hommeaux, E., Boneva, I., Kontokostas, D.: Validating RDF Data, Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 7. Morgan & Claypool Publishers LLC (sep 2017). https://doi.org/10.2200/s00786ed1v01y201707wbe016, https://doi.org/10.2200/s00786ed1v01y201707wbe016
46. Laukkanen, M., Viljanen, K., Apiola, M., Lindgren, P., Hyvönen, E.: Towards ontology-based yellow page services. In: Proceedings of WWW2004 Workshop, Application Design, Development, and Implementation Issues, New York (2004)

47. Leskinen, P., Hyvönen, E.: Linked open data service about historical Finnish academic people in 1640–1899. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming
48. Leskinen, P., Hyvönen, E., Tuominen, J.: Members of parliament in finland knowledge graph and its linked open data service (March 2021), aalto Univerisity, SeCo Group, submitted
49. Mäkelä, E., Hypén, K., Hyvönen, E.: BookSampo—lessons learned in creating a semantic portal for fiction literature. In: Proc. of ISWC-2011, Bonn, Germany. Springer (2011)
50. Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Reinsone, S., Skadiņa, I., Baklāne, A., Daugavietis, J. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96. CEUR Workshop Proceedings, CEUR-WS.org, Germany (2020), digital Humanities in the Nordic Countries, DHN2020 ; Conference date: 17-03-2020 Through 20-03-2020
51. Mäkelä, E., Ruotsalo, T., Hyvönen: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability **3**(1), 85–109 (2012)
52. March, S.T., Smith, G.F.: Design and natural science research on information technology. Decision Support Systems **15**(4), 251–266 (1995). https://doi.org/10.1016/0167-9236(94)00041-2
53. Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), 41–46 (2006)
54. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web – Interoperability, Usability, Applicability **11**(2), 255–335 (2020)
55. Miyakita, G., Leskinen, P., Hyvönen, E.: 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)
56. tuominen-et-al-avioand Nina Laurenne, J.T., Koho, M., Hyvönen, E.: The birds of the world ontology avio. In: The Semantic Web: ESWC 2013 Satellite Events. pp. 300–301. Springer (2013)
57. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. Journal of Management Information Systems **24**(3), 45–77 (2007). https://doi.org/10.2753/MIS0742-1222240302
58. Rantala, H., Ikkala, E., Jokipii, I., Hyvönen, E.: WarVictimSampo 1914–1922: a national war memorial on the semantic web for digital humanities research and applications. ACM Journal on Computing and Cultural Heritage (April 2021), https://seco.cs.aalto.fi/publications/2021/rantala-et-al-warvictimsampo-jochh-2021.pdf, in Press
59. Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web – Interoperability, Usability, Applicability **8**(3), 373–383 (2017). https://doi.org/10.3233/SW-150197
60. Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F.: Bioportal as a dataset of linked biomedical ontologies and terminologies in rdf. Semantic Web – Interoperability, Usability, Applicability **4**(3), 277–284 (2013). https://doi.org/10.3233/SW-2012-0086
61. Seppälä, K., Hyvönen, E.: Asiasanaston muuttaminen ontologiaksi. yleinen suomalainen ontologia esimerkkinä finnonto-hankkeen mallista (changing a keyword thesaurus into an ontology. general finnish ontology as an example of the finnonto model) (March 2014), https://www.doria.fi/handle/10024/96825
62. Sidoroff, T., Hyvönen, E.: Semantic e-goverment portals—a case study. In: Proceedings of the ISWC-2005 Workshop Semantic Web Case Studies and Best Practices for eBusi-

ness SWCASE05 (Nov 2005), http://www.seco.hut.fi/publications/2005/sidoroff-hyvonen-semantic-e-government-2005.pdf
63. Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., Mela, M.L., Hyvönen, E.: Plenary debates of the Parliament of Finland as linked open data and in Parla-CLARIN markup (March 2021), aalto Univerisity, SeCo Group, submitted
64. Staab, S., Studer, R. (eds.): Handbook on Ontologies (2nd Edition). Springer (2009)
65. Suominen, O., Hyvönen, E.: Improving the quality of SKOS vocabularies with Skosify. In: Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012). Springer (2012)
66. Suominen, O., Hyvönen, E., Viljanen, K., Hukka, E.: HealthFinland – a national semantic publishing network and portal for health information. Journal of Web Semantics **7**(4), 287–297 (2009)
67. Suominen, O., Pessala, S., Tuominen, J., Lappalainen, M., Nykyri, S., Ylikotila, H., Frosterus, M., Hyvönen, E.: Deploying national ontology services: From onki to finto. In: Proceedings of the Industry Track at the International Semantic Web Conference 2014. CEUR Workshop Proceedings (2014), http://www.ceur-ws.org/Vol-1383, vol 1383
68. Tamper, M., Oksanen, A., Tuominen, J., Hietanen, A., Hyvönen, E.: Automatic annotation service: Utilizing a named entity linking tool in legal domain (2019), submitted article under evaluation
69. Tunkelang, D.: Faceted search. Morgan & Claypool Publishers, CA, USA (2009)
70. Tuominen, J., Frosterus, M., Viljanen, K., Hyvönen, E.: 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)
71. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011). Springer (June 2011). https://doi.org/10.1007/978-3-642-21064-8_18
72. Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., Hotson, H.: Reassembling the republic of letters – a linked data approach. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 76–88. CEUR Workshop Proceedings, vol. 2084 (March 2018), http://www.ceur-ws.org/Vol-2084/paper6.pdf
73. Tzitzikas, Y., Manolis, N., Papadakos, P.: Faceted exploration of RDF/S datasets: a survey. Journal of Intelligent Information Systems **48**(2), 329–364 (2017)
74. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the ESWC 2009, Heraklion, Greece. pp. 781–795. Springer (2009)
75. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A survey. Semantic Web – Interoperability, Usability, Applicability **7**(1), 63–93 (2016). https://doi.org/10.3233/SW-150175

Table 3. A selection of Sampo portals and LOD services for Digital Humanities

#	Portal	Year	Domain	# Users	# Triples	Primary data owners
1	MuseumFinland	2004	Artefact collections	39 000	211 000	National Museums, City Museums of Espoo and Lahti, Finland
2	CultureSampo	2008	Finnish culture	107 000	11M	Memory organizations and the Web, ca 30 data sources
3	BookSampo	2011–	Fiction literature	2M/year	4,36M[a]	Public libraries in Finland (Kirjastot.fi)
4	WarSampo	2015–2019	World War II	740 000	14M	National Archives, Defense Forces, and others, Finland
5	Norssit Alumni	2017	Person registry	unknown	469 000	Norssi High School alumni organization Vanhat Norssit
6	U.S. Legislator Prosopographer	2018	Parliamentary data	unknown	830 000	U. S. Congress Legislator data[b]
7	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
8	BiographySampo	2019	Biographies	50 000	5,56M	Finnish Literature Society
9	WarVictimSampo 1914–1922	2019	Military history	29 000	9,96M	National Archives of Finland
10	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)
11	AcademySampo	2021	Finnish Academics	2100	6,55M	University of Helsinki and National Archives, Finland
12	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.