Creating and Using a Linked Open Ontology and Data Infrastructure for Digital Humanities in Finland: Lessons Learned 2003–2023

Eero Hyvönen[1,2]

[1] Aalto University, Department of Computer Science,
Semantic Computing Research Group (SeCo), Finland
[2] University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland
http://seco.cs.aalto.fi/u/eahyvone/

Abstract. This paper reports experiences in creating a national ontology and Linked (Open) Data (LOD) infrastructure for Digital Humanities in Finland (2003–2023) (LODI4DH), including centralized ontology and data services and tooling for creating applications. The LODI4DH infrastructure has been used in practise for creating a series of over twenty LOD services and portals in use. The portals have have attracted millions of users in total suggesting feasibility of the proposed model. This line of research and development is unique due to its systematic national level nature and long time span of twenty years.

Keywords: Semantic Web Linked Data Ontologies Web Services Infrastructures Portals Digital Humanities

1 Why Intrastructure for Digital Humanities?

The development of the Semantic Web (SW) was boosted by the seminal article in Scientific American in May 2011 by Tim Berners-Lee et al. [1]. After a few months the conference "Semantic Web (SW) Kick-off in Finland" was organized [2]. This event initiated SW research in Finland and the first in-use application "Promoottori" appeared soon for publishing conferment tradition data of the University of Helsinki [3,4].

In 2004, a demonstrator for a national platform for aggregating and publishing museum collections on the Web was published: *MuseumFinland – Finnish museums on the Semantic Web*[3] [5]. The key idea of this system was to create a collaborative publishing platform for collection data for museums that would enrich their data mutually by data linking and reasoning. For the end users, the system could provide enhanced functionalities, such as semantic search and intelligent browsing based on the larger aggregated and harmonized dataset from several museums. This kind of service could not be created using federated search then commonly used, but the heterogenous, distributed data from different museums had to be harmonized into a global RDF graph[4].

When creating MuseumFinland two major challenges for this kind of collaborative publishing model of Cultural Heritage (CH) content were identified:

[3] Project: https://seco.cs.aalto.fi/applications/museumfinland/; portal: http://museosuomi.fi

[4] https://www.w3.org/RDF/

1. *Interoperability problem*. The first key problem is that Cultural Heritage contents in museums, libraries, archives, and galleries are richly interlinked and heterogenous, not only on a national scale but across the borders of different countries and cultures. We all share common history and cultural background. The cultural content includes texts, speech, images, 2D and 3D models, movies, and animations represented in incompatible formats using different national data models and using different natural languages.
2. *Data production coordination problem*. The second key problem is that cultural heritage content creation is distributed in different kind of organizations, including not only memory organizations, but also media companies, land survey organizations, communities producing data, such as Wikipedias an Linked Data clouds, and citizens participating in citizen science initiatives. When parties create content indendently from each other, the result is an avalanche of data silos and web services that do not speak with each other.

A solution to both problems is to create a shared ontology infrastructure to be used by different stakeholders when creating their metadata. The problem of heterogeneity can be approached by using shared data models and by using shared knowledge organization systems (KOS) (vocabularies, thesauri, classifications, gazetteers) when populating the data models. The coordination problem can be approach by creating and standardizing collaboratively the data models and KOS used, and by sharing them through centralized ontology services. [6]

The work on creating a national level SW infrastructure in Finland stared by a series of projects called "FinnONTO" (2003–2012)[5], followed by "Linked Data Finland" projects (2012–2014) with a focus on Linked Open Data (LOD) services and projects on applying the infrastructure [7,8]. Much of this infrastructure work has focused on the CH domain coined as "Linked Open Data Infrastructure for Digital Humanities in Finland" (LODI4DH). This paper gives an overview of LODI4DH reporting on lessons learned when creating and using the LOD, ontology services, LOD services, and when developing practical applications on top of the infrastructure.

2 Seven Components of a Semantic Web Infrastructure

A system like LODI4DH should be based on widely used standards. In our case the obvious choice was to use the domain agnostic W3C *Web Standards and Best Practices* of publishing Linked Data[6] [9]. The FAIR principles[7], compatible with the linked data principles[8] and best practices[9] of the W3C, were employed for creating Findable, Accessible, Interoperable, and Re-usable data.

The key components of LODI4DH include the following seven components:

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

[6] Best Practices for Publishing Linked Data https://www.w3.org/TR/ld-bp/

[7] The FAIR principles: https://www.go-fair.org/fair-principles/

[8] Linked data design issues: https://www.w3.org/DesignIssues/LinkedData.html

[9] Data on the Web Best Practices: https://www.w3.org/TR/dwbp/

1. *Metadata models*. Shared models for metadata [10] are needed for representing knowledge of different application domains. In our case, we have used both document centric models, such as Dublin Core (DC), as well as more foundational ontological models for data harmonization, such as CIDOC CRM and the FRBR family of models for literary works [11]. Both approaches may be needed even for the same data in one system. For example, in the WarSampo system [12,13], death records are represented as documents using a DC-based model for presenting, querying, and maintaining the data, and as events, based on CIDOC CRM. Using the latter model, events such as birth, getting wounded and killed, could be integrated in the biographical event timelines of the soldiers' lives for interoperability.
2. *Domain Ontologies*. Shared domain ontologies are needed for populating the metadata models by resources taken from shared KOS for interoperability. In this paper, the term domain ontology refers to typically hierarchical, thesaurus-like knowledge organization systems whose concepts are used to populate property values of (meta)data models.
3. *Domain ontology services* The ontologies should be made openly available and easy to access for interoperability and re-use, based on shared ontology services; cf. [14,15] for a survey of such systems.
4. *Data services* In the same vein, data services for publishing LD datasets and their data models, preferably using, e.g., open Creative Commons licenses, are needed for making re-use of data possible and easy.
5. *Applications* Also Applications of Linked Data are part of the infrastructure connecting the system to its end users.
6. *Software Tools* 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 [16]. Also tools for data publishing and analysis are needed, as well as tooling for developing new applications for the end users.
7. *Human Infrastructure* For developing, maintaining, and using the infrastructure in a sustainable way a Human Infrastructure is needed, too. This involves, e.g., educating people about the technology[10], introducing SW courses in university curricula, and production of documentations and learning materials for the community using national languages.

When developing the Finnish SW infrastructure, applications that test and demonstrate its usability were constantly developed. This work has evolved into a set of principles for developing LOD services and semantic portals on top of them, called the *Sampo Model*[11] [8]. In this model, shared ontology services, data services, and tools for user interface design have turned out to be crucial for the practical implementation work. In

[10] See, e.g., the open self-study video lecture course "Linked Data Technologies for Cultural Heritage and Digital Humanities: Introducing the Semantic Web in Video Lectures" at https://seco.cs.aalto.fi/teaching/sw-introduction/.

[11] 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.

the following, the work on LODI4DH is overviewed from this application development and deployment point of view.

3 Applying the Semantic Web Infrastructure: Sampo Model

The Sampo model is an attempt to formulate a set of re-usable design principles or guidelines for creating LOD services and semantic portals, especially for Cultural Heritage applications and Digital Humanities research [17]. Based on six principles listed in Table 1, the model is a kind of consolidated approach[12] for creating LOD services and semantic portals.

Table 1. Sampo Model Principles P1–P6 [8]

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

The Sampo Model is an informal collection of principles for LOD publishing and designing semantic portals. Principles P1–P3 can be seen as a foundation for developing data services; P4–P6 are related to creating semantic portals. The model is based on the idea of collaborative content creation (P1). The data is aggregated from local data silos into a global service, based on a shared ontology infrastructure (P2). The 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. The model argues for the idea of separating the underlying Linked Data service *completely* from the user interface via a SPARQL API (P3). This arguable simplifies the portal architecture and the data service can be opened for data analysis research. For example, YASGUI[13] [19] editor for SPARQL querying and visualizing data can be used, or Python scripting in Google Colab[14] and Jupyter notebooks[15] [20].

The general idea of principles P4–P6 is to "standardize" the UI logic so that the portals are easier to use for the end users and for the programmers to develop [21]. Principle P4 articulates the idea of providing different thematic *application perspectives* by re-using the data service. The application perspectives can be provided on the landing page of the Sampo portal system or be completely separate applications by third parties. According to P5 the application perspectives can be used by a two-step cycle

[12] Something that the field of the SW is arguably still largely missing [18]

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

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

[15] https://jupyter.org

for research: First, the focus of interest, the target group, is filtered out using faceted semantic search [22,23,24]. Second, the target group is visualized or analyzed by using ready-to-use data analytic tools of the application perspectives. Finally, the Sampo model aims not only at data publishing with search and data exploration [25] but also to data analysis and knowledge discovery with seamlessly integrated tooling for finding, analysing, and even solving research problems in interactive ways (P6) [26].

The Sampo model has evolved gradually in 2002–2023 via lessons learned in developing a series of semantic portals and LOD services, starting from **MuseumFinland – Finnish Museums on the Semantic Web**[16] (online since 2004) [5], **CultureSampo – Finnish Culture on the Semantic Web 2.0**[17] (online since 2009) [27,28], and **BookSampo**[18] (online since 2011 with some 1.6 million annual users today) [29]. They demonstrated how CH content of dozens of different kinds, both tangible and intangible CH content, can enrich each other. **WarSampo – Finnish World War II on the Semantic Web**[19] (online since 2015 with several new perspectives published in 2016–2019) [12] is a popular Finnish service that has had thus far over million users. A key idea in WarSampo is to reassemble the life stories of the World War II soldiers based on data linking from different data sources. This biographical and prosopographical idea was a source of inspiration for several later biographical applications, including **BiographySampo – Biographies on the Semantic Web**[20] (online since 2018) [30], **Norssit Alumni** [31], **U.S. Congress Prosopographer** [32], and **AcademySampo**[21] (online since 2021) [33]. **NameSampo** [34] publishes data about over 2 million place names and places in Finland with old maps. The NameSampo project developed, based on the SPARQL Faceter tool [35] used in many earlier Sampos, the first version of the Sampo-UI framework [21] that has been used after this in all Sampos. It supports implementation of principles P4–P6 from an UI point of view. Sampo-UI has been re-used, e.g., in the portal **Mapping Manuscript Migrations (MMM)**[22] (online since 2020) [36,37] based on metadata about some 220 000 pre-modern manuscripts from the University of Oxford (U.K.), Schoenberg Institute (U.S.), and IRHT (France), in **FindSampo**[23] [38] (online since 2021) for supporting archaeology from a citizen science and metal detectorists' perspectives. **LetterSampo**[24] [39] is based on early modern epistolary metadata aggregated in the Early Modern Letters Online (EMLO) service[25] at the Oxford University, the CKCC corpus underlying ePistolarium[26] of the Huygens Institute in the Netherlands, and correspSearch[27] service of the Berlin-Brandenburg Academy of Sci-

[16] This application at https://museosuomi.fi got the Semantic Web Challenge Award at the ISWC 2004 conference.

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

[18] https://kirjasampo.fi

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

[20] https://seco.cs.aalto.fi/projects/biografiasampo/en/

[21] https://seco.cs.aalto.fi/projects/akatemiasampo/en/

[22] https://seco.cs.aalto.fi/projects/mmm/

[23] https://seco.cs.aalto.fi/projects/sualt/

[24] https://seco.cs.aalto.fi/projects/rrl/

[25] http://emlo.bodleian.ox.ac.uk

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

[27] https://correspsearch.net

ences. During the spring of 2023 two new Sampos of particular societal impact were released: **LawSampo**[28] [40] publishes Finnish legislation and case law based on data from the Ministry of Justice in Finland. **ParliamentSampo**[29] [41] publishes LOD of the Parliament of Finland (1907–2023), nearly a million speeches interlinked with an ontology of the Parliament of Finland.

A key idea in developing LODI4DH is re-using the elements of the infrastructure and developing them further step-by-step in a systematic way when developing new applications. As for tooling the work, the Sampo-UI framework [21] has turned out to be very effective tool in developing the portal user interfaces, and it has been used also by some external developers. Natural language processing (NLP) techniques have been another important category of tools in later Sampos, such as LawSampo and ParliamentSampo, where lots of data have been available only in unstructured textual form. During our work, external NLP tools were re-used and new ones developed for named entity recognition (NER) and linking (NEL), for automatic annotation of keywords, and for topical classification of texts [42,43]. For LawSampo also a pseudonymization tool called Anoppi was created [44] as personal information in court decisions cannot be disclosed on the Web.

Data about all over 20 Sampo portals, including links, videos, publications, and further information are available on the Sampo portals homepage[30].

4 Domain Ontology Infrastructure

Sampo systems make used a cloud of domain ontologies available through ontology services. LODI4DH identifies several basic types of domain ontologies. Firstly, there are domain ontologies of classes. For example, the concept of “Novel”, “City”, “Bird”, or “War” are classes and particular novels and cities would be their instances. Here RDF(S) and OWL semantics can be used. Secondly, there are “*instance ontologies*” enumerating individuals of the classes. It makes often sense to separate class and instance ontologies, as the number of instances can be very large (e.g., cities in geograzetteers, novels and copies of them, etc.). Thirdly, there are SKOS-based ontologies that, from a semantic point of view, are used for representing thesauri, classifications and other knowledge organization systems. Here the class of `skos:Concept` is instantiated for representing, e.g., terms in a thesaurus or categories in a classification system.

A Cloud of Linked Ontologies A central goal of FinnONTO was to create an interlinked cloud of 16 national ontologies [45] based on existing thesauri that were already used in different areas of the society. The transformation process was more ambitious than just transforming the traditional standard thesaurus format [46] into an RDF-based model, such as SKOS[31]. The thesauri were developed semantically a bit forward, using the OntoClean methodology [47] and RDFS[32], in the following ways [48,49]: 1) Multiple meanings of thesauri terms were disambiguated and relocated in

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

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

[30] Sampo portals’ homepage: https://seco.cs.aalto.fi/applications/sampo/

[31] SKOS Reference: https://www.w3.org/TR/skos-reference/

[32] RDF Schema: https://www.w3.org/TR/rdf-schema/

`rdfs:subClassOf` hierarchies. For example, the concept of *child*, a unique concept in the underlying General Finnish Thesaurus YSA, can refer to the class of young people, to a family relation type, or a social class (superconcept of *street child*). 2) The thesauri that were transformed did not differentiate whether the standard Broader Term (BT) relation [46] means the 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 by the RDFS semantics, so that the hierarchies could be used for reasoning, e.g., in query expansion and when using faceted semantic search in applications [8].

The largest ontology YSO (27 200 concepts) shared lots of concepts with all other ontologies, in some cases more than 50%. This suggested that the ontologies should be linked together using YSO as the top ontology. This idea resulted in creating the Finnish linked ontology cloud called KOKO[33] where the top ontology concepts of YSO are refined by subconcepts of interlinked domain specific ontologies [48,45].

Ontology Services According to the FinnONTO vision, the ontologies should be served not only through human readable browser interfaces[34], but also as centrally managed national ontology services using APIs. In this way, common functionalities of the services, such as (semantic) autocompletion [50], URI fetching, and query expansion [51], can be shared on a national level, and everybody would get access to the up-to-date versions of the ontologies. This would be cost-efficient on a national level and gradually leads to better interoperability of the data catalogued in different organizations. Centralized services is needed especially for smaller organizations that do not have much expertise and resources for developing their own web services.

Lessons Learned A key problem to be solved in FinnONTO was that large cross-domain thesauri are difficult to maintain. 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 in transforming thesauri into ontologies and in maintaining the linked ontology cloud and the collaboration network [45]. The traditional thesauri semantics [46] were refined only a little using RDFS but already this was a handful of work, as thousands of terms in the thesauri had to be manually checked and refined [49]. These new challenges are now being tackled by the Finto collaboration network[35] coordinated by the National Library. The FinnONTO initiative pointed out that lots of redundant work had been done in developing the thesauri in Finland as they shared lots concepts with each other. Redundant work can be better eliminated in the new KOKO model. Tools such as MUTU [52] were developed to support the ontology alignment.

The idea of creating a “living laboratory” of ONKI ontology services [53,54] on the Web turned out to be important 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

[33] The current version of KOKO is available at the Finto.fi service: urlhttps://finto.fi/koko/fi/.

[34] Two human interfaces were created, ONKI.fi (https://onki.fi) for RDFS and ONKI Light for SKOS domain ontologies (https://light.onki.fi).

[35] https://www.kansalliskirjasto.fi/fi/content/finto-5

no return" was reached as the number of ONKI API users were already in hundreds. The motto for the FinnONTO work was 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 has already been done in cataloguing [55].

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 [56,57], and nomenclatures and terminologies of medicine [58], such as Medical Subject Headings MESH[36].

The FinnONTO ontologies were published first in 2008 using the ONKI.fi ontology server[53]. As a next step, the ONKI Light service[37] [59] was developed and deployed in 2014 [60] by the National Library of Finland as the national Finto.fi service[38]. ONKI Light finally evolved into the open source Skosmos tool[39] [61] in use in several other organizations in Finland and internationally[40]. ONKI Light was based on a SPARQL endpoint. The idea was to separate the data service fully from the user interface, and use only SPARQL to access the data. This idea turned later useful when developing the Sampo model and Sampo-UI tool for semantic portals. The Finto.fi service has grown into a popular national open 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. A national level paradigm change has taken place in Finland on using linked light-weight ontologies instead of thesauri.

5 Data Services: 7-star Linked Data Deployment Scheme

LODI4DH includes the Linked Data Finland service LDF.fi[41] [62] platform for publishing datasets and (re-)using them via web services. A key component in LOD publishing is the SPARQL endpoint, but the platform should also support other functions [9]. 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 the W3C Service Description recommendation[42]. 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,

[36] https://finto.fi/mesh/fi/
[37] https://seco.cs.aalto.fi/services/onkilight/
[38] Available at: https://finto.fi
[39] https://skosmos.org/
[40] For a list of international services, see https://www.kiwi.fi/display/Finto/Skosmos-ohjelmisto0
[41] https://ldf.fi
[42] http://www.w3.org/TR/sparql11-service-description/

documenting, inspecting, and validating the data. LDF.fi is used primarily for reading RDF data by SPARQL queries, not for writing, although also this could be done using the SPARQL endpoint.

Linked data publications on the SW are typically evaluated with the W3C “5-star” deployment scheme[43], 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[44]. 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 given to a data publication, if the publication includes some kind of evaluation that the data actually conforms to the provided schemas using, e.g., SHACL[45] or ShEx[46] [63]. The idea here is to encourage publishers to publish high quality data as data quality of LD is a severe issue on the SW. To extend the model even further, the 8th star could perhaps be given to data if it is shown to actually represent to the real world correctly.

Schemas can be documented automatically in LDF.fi for the human reader using a schema documentation generator, in our case using SpecGen[47] and LODE[48]. 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[49] was created that analyzes a given dataset from this perspective and creates an HTML document that lists, e.g., statistics of vobabulary usage and raises up issues detected 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[50] for storing the primary data and a Varnish Cache web application accelerator[51] for routing URIs, content negotiation, and caching. For deployment of applications with a data service (cf., e.g., the MMM system [36]) a microservice architecture with Docker containers[52] 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 in Finland.

Lessons learned The Linked Data Finland platform has turned out to be useful for data-analytic research purposes and in developing applications (cf. Section 3). LDf.fi has been used for publishing some 100 linked datasets. Many of them are in use in

[43] https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/
[44] Such as the Burj Al Arab in United Arab Emirates
[45] https://www.w3.org/TR/shacl/
[46] https://shex.io/
[47] https://bitbucket.org/wikier/specgen/wiki/Home
[48] https://essepuntato.it/lode/
[49] http://vocab.at
[50] https://jena.apache.org/documentation/fuseki2/
[51] https://varnish-cache.org
[52] https://www.docker.com

semantic portal applications 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-based version of the ca. 800 000 official Finnish geographical places based on data of the National Survey.

6 Discussion

This paper presented LODI4DH infrastructure in use in Finland for developing LOD services and applications for DH. Our general strategy has been to develop useful proof-of-concept prototypes and to publish them openly on the Web for everyone to use. The data owners and stake holders, such as memory organizations, saw this as an opportunity to develop their own systems, started to use the services and applications, and in many cases the point of no return has been reached. Our work first focused on domain ontologies and ontology services (ONKI.fi and Finto.fi) and then on Linked Data and data model services. To test and demontrate LODI4DH a series of "Sampo" LOD services and portals have been developed and are in use. As new ontologies and applications with new datasets are developed, the open LOD already available in the infrastructure, say ontologies of places and historical people, can be reused and refined gradually better and better. This applies also to the open source tools, such as the Sampo-UI framework.

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. 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 requires new kind of digital data literacy and source criticism[53] from the end user [64,65]. In spite of the challenges, enriching data carefully with semantics, in one way or another, is in my mind a way ahead towards creating a more and more intelligent Web in a cost-efficient way. In contrast using "black box" language model-based systems and deep machine learning, such as Chat GPT, the SW makes the data on the Web explicit, transparent, and well-defined, and the already structured curated data in databases can be utilized. This facilitates creation of explainable "white box" AI systems [26,66].

Acknowledgments Nearly 100 people[54] at the Semantic Computing Research Group (SeCo) have contributed to the work presented in this paper, funded by over 50 organizations in Finland and beyond. Our work is partly related to the Semantic Parliament (ParliamentSampo) project funded by the Academy of Finland, the EU project InTaVia:

[53] https://ranke2.uni.lu/define-dsc/#%20,%20Universit%C3%A9%20du%20Luxembourg

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

In/Tangible European Heritage[55], and the EU COST action Nexus Linguarum[56] on linguistic data science. Thanks to the Finnish Cultural Foundation for an Eminentia Grant for reflecting the work presented. CSC – IT Center for Science has provided computational resources for our projects.

References

1. T. Berners-Lee, Design Issues: Linked Data, 2006, https://www.w3.org/DesignIssues/LinkedData.html. https://www.w3.org/DesignIssues/LinkedData.html.
2. E. Hyvönen (ed.), Semantic Web Kick-Off in Finland – Vision, Technologies, Research, and Applications, in *HIIT Publications 2002-01*, Helsinki Institute for Information Technology, 2002. https://seco.cs.aalto.fi/publications/2002/hyvonen-semantic-web-kick-off-2002.pdf.
3. E. Hyvönen, A. Styrman and S. Saarela, Ontology-Based Image Retrieval, in: *Towards the semantic web and web services, Proceedings of XML Finland 2002 Conference*, 2002, pp. 15–27.
4. E. Hyvönen, S. Saarela and K. Viljanen, Application of Ontology Techniques to View-Based Semantic Search and Browsing, in: *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, Springer, 2004, pp. 92–106. doi:10.1007/978-3-540-25956-5_7.
5. E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila and S. Kettula, MuseumFinland—Finnish Museums on the Semantic Web, *Journal of Web Semantics* **3**(2) (2005), 224–241.
6. E. Hyvönen, *Publishing and using cultural heritage linked data on the Semantic Web*, Morgan & Claypool, Palo Alto, California, 2012.
7. E. Hyvönen, How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web, *Semantic Web – Interoperability, Usability, Applicability* (2023), Forth-coming.
8. E. Hyvönen, Digital Humanities on the Semantic Web: Sampo Model and Portal Series, *Semantic Web – Interoperability, Usability, Applicability* **14**(4) (2022), 729–744. doi:10.3233/SW-190386.
9. T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space (1st edition)*, Morgan & Claypool, Palo Alto, California, 2011. http://linkeddatabook.com/editions/1.0/.
10. M. Zeng and J. Qin, *Metadata, Third Edition*, ALA Neal-Schuman, Chicago, 2022. ISBN ISBN 978-0-8389-4875-0.
11. P. Riva, M. Doerr and M. Zumer, FRBRoo: Enabling a Common View of Information from Memory Institutions, *International Cataloguing and Bibliographic Control (ICBC)* **38**(2) (2009).
12. E. Hyvönen, E. Heino, P. Leskinen, E. Ikkala, M. Koho, M. Tamper, J. Tuominen and E. Mäkelä, WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History, in: *The Semantic Web – Latest Advances and New Domains (ESWC 2016)*, Springer, 2016, pp. 758–773. doi:978-3-319-34129-3_46.
13. M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen and E. Hyvönen, WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data, *Semantic Web – Interoperability, Usability, Applicability* **12**(2) (2021), 265–278. doi:10.3233/SW-200392.
14. M. d'Aquin and N.F. Noy, Where to Publish and Find Ontologies? A Survey of Ontology Libraries, *Web Semantics: Science, Services and Agents on the World Wide Web* **11** (2012), 96–111.

[55] https://intavia.eu/

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

15. D. Naskar and B. Dutta, Ontology And Ontology Libraries: A Study From An Ontofier And An Ontologist Perspective, in: *ETD 2016 "Data and Dissertations". 19th International Symposium on Electronic Theses and Dissertations July, 11–13, 2016, Lille, France*, 2016, pp. 1–12, Lille, France. https://etd2016.sciencesconf.org/92726.html.
16. J.L. Martinez-Rodriguez, A. Hogan and I. Lopez-Arevalo, Information Extraction Meets the Semantic Web: A Survey, *Semantic Web – Interoperability, Usability, Applicability* **11**(2) (2020), 255–335. doi:10.3233/SW-180333.
17. E. Gardiner and R.G. Musto, *The Digital Humanities: A Primer for Students and Scholars*, Cambridge University Press, New York, NY, USA, 2015, https://doi.org/10.1017/CBO9781139003865.
18. P. Hitzler, A Review of the Semantic Web Field, *Commun. ACM* **64**(2) (2021), 76–83–. doi:10.1145/3397512.
19. L. Rietveld and R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web – Interoperability, Usability, Applicability* **8**(3) (2017), 373–383. doi:10.3233/SW-150197.
20. M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen and E. Hyvönen, Automatic Annotation Service: Utilizing a Named Entity Linking Tool in Legal Domain, in: *The Semantic Web: ESWC 2020 Satellite Events*, Springer, 2019, pp. 208–213. doi:978-3-030-62327-2_36.
21. E. Ikkala, E. Hyvönen, H. Rantala and M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, *Semantic Web – Interoperability, Usability, Applicability* **13**(1) (2022), 69–84. doi:10.3233/SW-210428.
22. E. Hyvönen, S. Saarela and K. Viljanen, Application of ontology-based techniques to view-based semantic search and browsing, in: *Proceedings of the First European Semantic Web Symposium*, Springer, 2004.
23. D. Tunkelang, *Faceted search*, Morgan & Claypool, Palo Alto, California, 2009. doi:10.2200/S00190ED1V01Y200904ICR005.
24. Y. Tzitzikas, N. Manolis and P. Papadakos, Faceted exploration of RDF/S datasets: a survey, *Journal of Intelligent Information Systems* **48**(2) (2017), 329–364. doi:10.1007/s10844-016-0413-8.
25. G. Marchionini, Exploratory search: from finding to understanding, *Communications of the ACM* **49**(4) (2006), 41–46. doi:10.1145/1121949.11219790.
26. E. Hyvönen, 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) (2020), 187–193. doi:10.3233/SW-190386.
27. E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio and K. Nyberg, CultureSampo – Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user, in: *Museums and the Web 2009*, Archives & Museum Informatics, Toronto, 2009.
28. E. Mäkelä, T. Ruotsalo and Hyvönen, How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo, *Semantic Web – Interoperability, Usability, Applicability* **3**(1) (2012), 85–109. doi:10.3233/SW-2012-0049.
29. E. Mäkelä, K. Hypén and E. Hyvönen, BookSampo—Lessons Learned in Creating a Semantic Portal for Fiction Literature, in: *The Semantic Web – ISCW 2011*, Springer, 2011, pp. 173–188. doi:10.1007/978-3-642-25093-4_120.
30. E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen and K. Keravuori, BiographySampo – Publishing and Enriching Biographies on the Semantic Web for Digital Humanities Research, in: *The Semantic Web. ESWC 2019*, Springer, 2019, pp. 574–589. doi:978-3-030-21348-0_37.
31. E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen and L. Sirola, Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Se-

mantic Web, in: *Proceedings, Language, Technology and Knowledge (LDK 2017)*, Springer, 2017, pp. 113–119. doi:978-3-319-59888-8_9.

32. G. Miyakita, P. Leskinen and E. Hyvönen, Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators, in: *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus*, Springer, 2018, pp. 150–162. doi:10.1007/978-3-030-01765-1_18.
33. P. Leskinen and E. Hyvönen, Linked Open Data Service about Historical Finnish Academic People in 1640–1899, in: *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, Vol. 2612, 2020, pp. 284–292. http://ceur-ws.org/Vol-2612/short14.pdf.
34. E. Ikkala, J. Tuominen, J. Raunamaa, T. Aalto, T. Ainiala, H. Uusitalo and E. Hyvönen, NameSampo: A Linked Open Data Infrastructure and Workbench for Toponomastic Research, in: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities*, GeoHumanities'18, ACM, New York, NY, USA, 2018, pp. 2:1–2:9. ISBN ISBN 978-1-4503-6032-6. doi:10.1145/3282933.3282936.
35. M. Koho, E. Heino and E. Hyvönen, 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, Vol. 1615. http://www.ceur-ws.org/Vol-1615.
36. E. Hyvönen, E. Ikkala, M. Koho, J. Tuominen, T. Burrows, L. Ransom and H. Wijsman, Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research, in: *The Semantic Web – ISWC 2021*, Springer, 2021, pp. 615–630. ISBN ISBN 978-3-030-88360-7. doi:10.1007/978-3-030-88361-4_36.
37. M. Koho, T. Burrows, E. Hyvönen, E. Ikkala, K. Page, L. Ransom, J. Tuominen, D. Emery, M. Fraas, B. Heller, D. Lewis, A. Morrison, G. Porte, E. Thomson, A. Velios and H. Wijsman, Harmonizing and Publishing Heterogeneous Pre-Modern Manuscript Metadata as Linked Open Data, *Journal of the Association for Information Science and Technology (JASIST)* **73**(2) (2022), 240–257. doi:10.1002/asi.24499.
38. E. Hyvönen, H. Rantala, E. Ikkala, M. Koho, J. Tuominen, B. Anafi, S. Thomas, A. Wessman, E. Oksanen, V. Rohiola, J. Kuitunen and M. Ryyppö, Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework, *Antiquity, A Review of World Archaeology* **95**(382) (2021), E24. doi:10.15184/aqy.2021.87.
39. J. Tuominen, E. Mäkelä, E. Hyvönen, A. Bosse, M. Lewis and H. Hotson, Reassembling the Republic of Letters – A Linked Data Approach, in: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, CEUR Workshop Proceedings, Vol. 2084, 2018, pp. 76–88. http://www.ceur-ws.org/Vol-2084/paper6.pdf.
40. E. Hyvönen, M. Tamper, E. Ikkala, S. Sarsa, A. Oksanen, J. Tuominen and A. Hietanen, Publishing and Using Legislation and Case Law as Linked Open Data on the Semantic Web, in: *The Semantic Web: ESWC 2020 Satellite Events*, Lecture Notes in Computer Science, Vol. 12124, Springer, 2020, pp. 110–114. doi:10.1007/978-3-030-62327-2_19.
41. E. Hyvönen, L. Sinikallio, P. Leskinen, M.L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal and J. Kesäniemi, Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language, in: *Digital Parliamentary data in Action (DIPADA 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference*, CEUR WS Proceedings, Vol. 3133, 2022. https://ceur-ws.org/Vol-3133/paper05.pdf.
42. M. Tamper, P. Leskinen, E. Ikkala, A. Oksanen, E. Mäkelä, E. Heino, J. Tuominen, M. Koho and E. Hyvönen, AATOS – a Configurable Tool for Automatic Annotation, in: *Proceed-*

ings, Language, Data and Knowledge (LDK 2017), Springer, 2017. doi:10.1007/978-3-319-59888-8_24.

43. R. Leal, J. Kesäniemi, M. Koho and E. Hyvönen, Relevance Feedback Search Based on Automatic Annotation and Classification of Texts, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Open Access Series in Informatics (OASIcs), Vol. 93, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 18:1–18:15. doi:10.4230/OASIcs.LDK.2021.18.
44. A. Oksanen, M. Tamper, J. Tuominen, A. Hietanen and E. Hyvönen, Anoppi: A Pseudonymization Service for Finnish Court Documents, in: *Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference*, M. Araszkiewicz and V. Rodríguez-Doncel, eds, IOS Press, 2019, pp. 251–254. doi:10.3233/FAIA190335.
45. M. Frosterus, J. Tuominen, S. Pessala and E. Hyvönen, Linked Open Ontology cloud: managing a system of interlinked cross-domain light-weight ontologies, *International Journal of Metadata, Semantics and Ontologies* **10**(3) (2015), 189–201. http://dx.doi.org/10.1504/IJMSO.2015.073879.
46. J. Aitchison, A. Gilchrist and D. Bawden, *Thesaurus Construction and Use: A Practical Manual*, Aslib IMI, 2000.
47. N. Guarino and C. Welty, Evaluating Ontological Decisions with OntoClean, *Communications of the ACM* **45**(2) (2002), 61–65. doi:10.1145/503124.503150.
48. E. Hyvönen, K. Viljanen, J. Tuominen and K. Seppälä, Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach, in: *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008*, Springer, 2008, pp. 95–109. doi:10.1007/978-3-540-68234-9_10.
49. K. Seppälä and E. Hyvönen, 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), National Library, Plans, Reports, Guides, 2014. https://www.doria.fi/handle/10024/96825.
50. E. Hyvönen and E. Mäkelä, Semantic Autocompletion, in: *The Semantic Web – ASWC 2006. First Asian Semantic Web Conference, Proceedings*, Springer, 2006, pp. 739–751. doi:10.1007/11836025_72.
51. J. Tuominen, T. Kauppinen, K. Viljanen and E. Hyvönen, Ontology-Based Query Expansion Widget for Information Retrieval, in: *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009)*, CEUR Workshop Proceedings, Vol. 449, 2009. http://ceur-ws.org/Vol-449/.
52. S. Pessala, K. Seppälä, O. Suominen, M. Frosterus, J. Tuominen and E. Hyvönen, MUTU: An Analysis Tool for Maintaining a System of Hierarchically Linked Ontologies, in: *Proceedings of the Workshop on Ontologies come of Age Workshop (ISWC 2011)*, 2011. https://seco.cs.aalto.fi/publications/2011/pessala-et-al-mutu-2011.pdf.
53. K. Viljanen, J. Tuominen and E. Hyvönen, Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI, in: *The Semantic Web: Research and Applications (Proceedings of ESWC 2009)*, Springer, 2009, pp. 781–795. doi:10.1007/978-3-642-02121-3_57.
54. J. Tuominen, M. Frosterus, K. Viljanen and E. Hyvönen, ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services, in: *The Semantic Web: Research and Applications. ESWC 2009*, Springer, 2009. doi:10.1007/978-3-642-02121-3_56.
55. E. Hyvönen, Preventing interoperability problems instead of solving them, *Semantic Web – Interoperability, Usability, Applicability* **1**(1–2) (2010), 33–37. doi:10.3233/SW-2010-0014.
56. J. Tuominen, N. Laurenne, M. Koho and E. Hyvönen, The Birds of the World Ontology AVIO, in: *The Semantic Web: ESWC 2013 Satellite Events*, Springer, 2013, pp. 300–301. doi:10.1007/978-3-642-41242-4_51.

57. J. Tuominen, N. Laurenne and E. Hyvönen, 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, 2011. doi:10.1007/978-3-642-21064-8_18.
58. O. Suominen, E. Hyvönen, K. Viljanen and E. Hukka, HealthFinland – a National Semantic Publishing Network and Portal for Health Information, *Journal of Web Semantics* **7**(4) (2009), 287–297. doi:10.2139/ssrn.3199436.
59. O. Suominen, A. Johansson, H. Ylikotila, J. Tuominen and E. Hyvönen, Vocabulary Services Based on SPARQL Endpoints: ONKI Light on SPARQL, in: *Poster proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2012)*, 2012. https://seco.cs.aalto.fi/publications/2012/suominen-et-al-onkilight-2012.pdf.
60. O. Suominen, S. Pessala, J. Tuominen, M. Lappalainen, S. Nykyri, H. Ylikotila, M. Frosterus and E. Hyvönen, 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, Vol 1383. ISSN 1613-0073. http://www.ceur-ws.org/Vol-1383.
61. O. Suominen, H. Ylikotila, S. Pessala, M. Lappalainen, M. Frosterus, J. Tuominen, T. Baker, C. Caracciolo and A. Retterath, Publishing SKOS vocabularies with Skosmos (2015), National Library of Finland, Manuscript. https://skosmos.org/publishing-skos-vocabularies-with-skosmos.pdf.
62. E. Hyvönen, J. Tuominen, M. Alonen and E. Mäkelä, Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets, in: *ESWC 2014: The Semantic Web: ESWC 2014 Satellite Events*, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.
63. J.E. Labra Gayo, E. Prud'hommeaux, I. Boneva and D. Kontokostas, *Validating RDF Data*, Synthesis Lectures on the Semantic Web: Theory and Technology, Vol. 7, Morgan & Claypool Publishers LLC, 2017, pp. 1–328. doi:10.2200/s00786ed1v01y201707wbe016.
64. T. Koltay, Data literacy for researchers and data librarians, *Journal of Librarianship and Information Science* **49**(1) (2015), 3–14. doi:10.1177/0961000615616450.
65. E. Mäkelä, K. Lagus, L. Lahti, T. Säily, M. Tolonen, M. Hämäläinen, S. Kaislaniemi and T. Nevalainen, Wrangling with non-standard data, in: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, CEUR Workshop Proceedings, Vol. 2612, 2020, pp. 81–96. https://ceur-ws.org/Vol-2612/paper6.pdf.
66. G. Vilone and L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, *Information Fusion* **76** (2021), 89–106. doi:10.1016/j.inffus.2021.05.009.