Linked Open Data Infrastructure for Digital Humanities in Finland

Eero Hyvönen

University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), and
Aalto University, Semantic Computing Research Group (SeCo)
http://heldig.fi, http://seco.cs.aalto.fi
eero.hyvonen@aalto.fi

Abstract. This paper presents and overviews *Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)*, a joint initiative of Aalto University, Department of Computer Science, and University of Helsinki (UH), Helsinki Centre for Digital Humanities (HELDIG), for creating a national data infrastructure and Linked Data services for open science. The data and services enable publication and utilization of datasets for data-intensive Digital Humanities (DH) research in structured, standardized formats via open interfaces. LODI4DH is based on a large national collaboration network and software created during a long line of national projects in DH between UH and Aalto since 2003. This work has created several in-use infrastructure prototypes, such as the ONKI and Finto ontology service now at the National Library of Finland, the Linked Data Finland data publishing platform LDF.fi, and the "Sampo" series of semantic portals testing and demonstrating the usability of the approach in practical applications. Thus far, these systems have had millions of end-users on the Web suggesting a high potential of utilizing the technology and Linked Data infrastructure in DH.

Keywords: Semantic Web, Linked Data, Infrastructure, Digital Humanities

1 A National Vision since 2003

Data, the oil of the digital world, is typically interlinked in content, published in different formats and languages, and is distributed in different services across countries. To create intelligent web services based on such heterogeneously distributed data sources, e.g., collection data in different cultural heritage organizations, the data has to be made mutually interoperable and machine understandable. Furthermore, lots of duplicate work can be eliminated if the data is not only interoperable, but also findable, accessible, and re-usable between the data publishers and users. To satisfy these FAIR principles[1], a shared data infrastructure is needed on a national level [5], interlinked with the international Semantic Web infrastructure and standards[2] using the Linked Data publishing principles [3].

[1] https://www.go-fair.org/fair-principles/
[2] https://www.w3.org/standards/semanticweb/

Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)[3] is a joint initiative of the Helsinki Centre for Digital Humanities (HELDIG)[4], the University of Helsinki (UH), and Aalto University Department of Computer Science, Semantic Computing Research Group (SeCo)[5]. The initiative aims to create living lab prototypes of centralized national Linked Data services for open science. The services enable publication and utilization of datasets for data-intensive Digital Humanities (DH) research and application in structured, standardized Semantic Web formats via open interfaces, such as SPARQL endpoints. Major components of LODI4DH include domain ontologies for data linking (historical places and maps, persons, events, keyword concepts, and times), harmonizing metadata models (extending, e.g., the CIDOC CRM ontology), publishing core datasets for re-use, various Linked Data services for developers, and online learning materials about Linked Data technologies.

LODI4DH is based on a large collaboration network and software infrastructure created during a long line of national projects in DH between UH and Aalto University, such as the FinnONTO project series (2003–2012)[6] on creating a national ontology infrastructure in Finland [5], the Linked Data Finland project forcusing on datasets (2012–2014)[7], and Linked Open Data Science[8]. These project collaborations between some 50 organizations have resulted in several in-use infrastructure prototypes and services, such as the ONKI ontology service, deployed as the Finto ontology service[9] by the National Library of Finland, and the "7-star" Linked Data Finland model and online service LDF[10] [10].

2 Ontology and Data Services in Use

ONKI/Finto and Linked Data Finland services have already had a wide user base demonstrating the need for the LODI4DH infrastructure. Applications based on them have also made their way from academic research into real use, especially the "Sampo" series of semantic portals [6] for Digital Humanities. For example, the system "BookSampo – Finnish Fiction Literature on the Semantic Web", based originally on LDF.fi and maintained now by the Finnish public libraries (Kirjastot.fi), had 2 million visitors in 2018, the semantic portal "WarSampo – Finnish World War II on the Semantic Web" (2015) has had 570 000 distinct users, and there are tens of thousands of monthly users in the Finto service. Many museums in Finland, e.g., Espoo City Museum, AKSELI Consortium of eight museums, and the new national MuseumPlus cataloging system, make use of the FinnONTO ontologies and the Finto service.

In addition to the Finnish projects, there have been several joint research projects and collaborations with foreign research organizations, such as University of Oxford,

[3] https://seco.cs.aalto.fi/projects/lodi4dh/
[4] http://heldig.fi
[5] http://seco.cs.aalto.fi
[6] https://seco.cs.aalto.fi/projects/finnonto/
[7] https://seco.cs.aalto.fi/projects/ldf/
[8] https://seco.cs.aalto.fi/projects/lodsci/
[9] http://finto.fi/en/
[10] http://ldf.fi

Stanford University, University of Colorado Boulder, and University of Pennsylvania, and Institut de recherche et d'histoire des textes (IRHT), where the Finnish Linked Data services (LDF.fi) for DH have been used. These include WW1LOD, a data service and semantic portal based on World War I data [13], a prototype [16] hosting data about ca. 150 000 letters for the Reassembling the Republic of Letters[11] initiative dealing with Early Modern correspondence data [4], and Mapping Manuscript Migrations[12] data service and portal [8] on medieval pre-modern manuscript data [1].

LODI4DH aims at harnessing all this work into sustainable services, and hopefully integrating the work as a component into the EU ERIC DARIAH infrastructure[13] in the future. LODI4DH infrastructure is open source, publishes open data (unless there are restrictions from the data owner side), and is free of charge for everyone to use. LODI4DH focuses on DH research infrastructures but the underlying Linked Data and Semantic Web technology can and has been utilized in other fields of research, too, extending the utilization potential of the infrastructure.

3 LODI4DH Components

Our research continues the work of the FinnONTO initiative on creating a shared ontology infrastucture [5]. Data from collaborating organizations is aggregated into shared open domain ontologies, including 1) historical places and maps, 2) historical persons, 3) events, 4) keyword concepts, and 5) times. These core ontologies, provided as web services, are used as "semantic glue" in data linking and fusion.

Historical Places and Maps As for historical places and maps, our work aims at developing the Finnish Ontology Service of Historical Places and Maps (Hipla.fi)[14] [7].

Historical Persons This work started already in FinnONTO, and has been revitalized in the context of building the National Semantic Biography of Finland, BiographySampo [9], AcademySampo [12] and related other biographical and prosopographical systems, based on Linked Data.[15]

Historical Events This line of research in LODI4DH builds upon our work on History on the Semantic Web, with applications such as WW1LOD, WarSampo – Finnish WW2 on the Semantic Web[16], and WarWictimSampo [14] on Finnish war history 1914–1922[17], and the Finnish History Ontology HISTO.[18]

[11] http://www.republicofletters.net/
[12] http://mappingmanuscriptmigrations.org/
[13] http://dariah.eu
[14] https://seco.cs.aalto.fi/projects/histoplaces/en/
[15] https://seco.cs.aalto.fi/projects/biographies/
[16] https://seco.cs.aalto.fi/projects/sotasampo/en
[17] https://seco.cs.aalto.fi/projects/sotasurmat-1914-1922/en/
[18] https://seco.cs.aalto.fi/projects/history/

Historical Keyword Concepts When developing ONKI and Finto, lots of Finnish keyword thesauri were converted and developed further into RDFS and SKOS ontologies, interlinked into a global linked data cloud called the KOKO [2], and published as ontology services. However, more work is needed here, for example, in areas such as archaeology, built environments, and law. For example, at moment we work on the AMMO ontology of thousands of historical Finnish professions[19] [11] interlinked with the international HISCO classification, and ontologies for describing archaeological finds[20], with interoperability to the AriadnePlus infrastructure[21] on an EU level.

Historical Times LODI4DH creates a time ontology for making references to historical times and periods of time, including names of time periods. Here results from international projects, such as PeriodO[22] can be utilized.

Harmonizing Metadata Models LODI4DH works on developing harmonizing metadata models for representing semantic data, such as Bio CRM [15] for extending CIDOC CRM to representing biographical and prosopographical data[23].

Core Datasets We also work on publishing and sharing interlinked core datasets, that are deemed to be useful in different research projects and applications. These datasets are expected to evolve into a kind of Finnish Linked Open Data Cloud. Work is going on, e.g., with the following datasets: Linked Open Name Archive, based on data about 2.7 million place names provided by the Institute for the Languages in Finland (Kotus); Semantic National Biography, based on over 13 100 biographies of prominent Finns edited by the Finnish Literature Society (SKS); WarSampo datasets related to WW2 history, provided by the National Archives of Finland, Defence Forces, and others; University of Helsinki Student Registry (1640–1899), provided by the University of Helsinki Archives; Semantic Finlex legislation and case law data, provided by the Ministry of Justice; Archaeological databases on the Finnish Heritage Agency; parliamentary discussions and other data from the Parliament of Finland.[24]

Linked Data Services As for the publishing platform, the "7-star" Linked Data Finland model and platform (LDF.fi) is used and developed further with additional services for DH data production, publishing, data analysis, validation, and visualization. LDF.fi extends Tim Berners-Lee's famous 5-star model[25] by two additional stars: the 6th start is given, if the dataset is published with the schemas it conforms to. The 7th start is given if an analysis of the quality of the data with respect to the schemas is provided, too [10].

[19] https://seco.cs.aalto.fi/ontologies/ammo/
[20] https://seco.cs.aalto.fi/projects/sualt/
[21] https://ariadne-infrastructure.eu/
[22] https://perio.do/en/
[23] https://seco.cs.aalto.fi/projects/biographies/
[24] For a full list of project homepages, see https://seco.cs.aalto.fi/projects/
[25] https://5stardata.info/en/

Learning Materials We also produce educational online materials, developing, e.g., the Linked Data School LinDa[26], for using Linked Data technology in DH research and application development.

4 Discussion: Sustainability

The end users of the LODI4DH applications have been both researchers and the public in the large, and the ontology services are used by professional catalogers in memory organizations, too. Our experiences suggest that building and maintaining a shared infrastructure is quite essential in developing applications for Digital Humanities effectively. In our own case studies, we have been able to reuse repeatedly and to develop further the FinnONTO ontologies, datasets, web services (e.g., map services), and software components (e.g., faceted search engines, language technology tools) in novel configurations and applications. Without sustainable reuse of the infrastructure, developing, e.g., the wide range of the Sampo portals [6] would have been impossible.

Deploying research prototypes into practical usage, however, can be time consuming. For example, it took ten years before our ONKI ontology service prototype was developed and deployed as the Finto service and got funded by the ministries as a national open service. The work faced not only technical challenges, but also issues of copyright of vocabularies and resistance of using ontologies and Linked Data in the first place. Our strategy for sustainability has been to develop living lab services, and see in practise if the end users get interested in using them, which would show that real add-on value has been created. If a "point of no return" can be reached, the problem of sustainability can be solved more easily. For example, at the moment hundreds of thousands of people have been using the WarSampo infrastructure (ontologies and data) and portal still maintained in a university environment, as the system is still used in our research. However, there are already negotiations on hosting and maintaining the system by external organizations. Pulling out the plug would be difficult at this point.

Acknowledgements Tens of people have been working in developing the ontologies and datasets of LOD4DH since 2003. Our data is based on databases of various memory and other organizations willing to open them for the public good. The research has been funded by ca. 50 organizations in Finland. The sites referred to in the footnotes contain full sets of publications online related to the systems, authored by the project members, as well as links to data, data services, and software. Thanks to CSC – IT Center for Science, Finland, for providing computational resources for LODI4DH.

References

1. Burrows, T., Hyvönen, E., Ransom, L., Wijsman, H.: Mapping manuscript migrations. Digging into data for the history and provenance of medieval and renaissance manuscripts. Manuscript Studies. A Journal of the Schoenberg Institute for Manuscript Studies **3**(1), 249–252 (2018), https://mss.pennpress.org/home/

[26] http://linda.seco.cs.aalto.fi/

2. 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
3. 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/
4. Hotson, H., Wallnig, T. (eds.): Reassembling the Republic of Letters in the Digital Age. Göttingen University Press (2019), https://doi.org/10.17875/gup2019-1146
5. 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–Verlag (2008)
6. Hyvönen, E.: "Sampo" model and semantic portals for Digital Humanities on the Semantic Web. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming
7. Hyvönen, E., Ikkala, E., Tuominen, J.: Linked data brokering service for historical places and maps. In: Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe). pp. 39–52. CEUR Workshop Proceedings (May 2016), http://ceur-ws.org/Vol-1608/, vol 1608
8. 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), http://www.ceur-ws.org/Vol-2364/
9. 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–Verlag (2019)
10. 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: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 226–230. Springer-Verlag (May 2014)
11. Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., Hyvönen, E.: AMMO Ontology of Finnish Historical Occupations. In: Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19). vol. 2375, pp. 91–96. CEUR Workshop Proceedings (June 2019), http://ceur-ws.org/Vol-2375/, vol 2375
12. 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
13. Mäkelä, E., Törnroos, J., Lindquist, T., Hyvönen, E.: WW1LOD: An application of CIDOC-CRM to World War 1 linked data. International Journal on Digital Libraries **18**(4), 333–343 (2017). https://doi.org/10.1007/s00799-016-0186-2
14. Rantala, H., Jokipii, I., Koho, M., Ikkala, E., Tuominen, J., Hyvönen, E.: Building a linked open data portal of war victims in Finland 1914–1922. In: Proc. of the Digital Humanities in the Nordic Countries (DHN 2020). CEUR WS Proceedings (2020), forth-coming
15. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). vol. 2119, pp. 59–66. CEUR Workshop Proceedings (2018), http://ceur-ws.org/Vol-2119/paper10.pdf
16. 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 (2018), http://www.ceur-ws.org/Vol-2084/paper6.pdf