How to Enhance Data Literacy of a Cloud of Heterogeneous Cultural Heritage Collections: Case SampoSampo – Connecting Everything to Everything Else

Henna Poikkimäki[1,*], Eero Hyvönen[1,2] and Petri Leskinen[1,2]

[1]*Aalto University, Espoo, Finland*
[2]*University of Helsinki, Helsinki, Finland*

Abstract

Computational analysis of cultural heritage (CH) data can be difficult due to distribution of the data into separate data sources, where different data models are used, inconsistent data may appear, formats used vary from scanned images to machine readable data, different natural languages are used, and different identifiers are used for resources. Aggregated web services for CH data have been built, such as the pan-European Europeana and the Sampo systems in Finland, to provide enriched global views to distributed local data sources. This paper argues that in such systems it is important not only to provide the aggregated data but also make its heterogeneity transparent to the end user; this is especially important if the aggregated data is used for data analyses. To facilitate this, we present a case study using the Linked Open Data (LOD) alignment service and web portal *SampoSampo – Connecting Everything to Everything Else*. It aligns entities such as people, organizations, and places in different collections in a way similar to the international VIAF.org mapping service for national library collections around the world. However, SampoSampo but is built on CH knowledge graphs (KGs) from different application domains, not only libraries, and includes also a semantic portal for end users. The challenge from a data analysis point of view in systems like these is that the data about entities in multiple Knowledge Graphs (KG) can be not only mutually complementary but also incomplete and conflicting and data provenience is important, i.e., how the data has been enriched. This paper demonstrates how such aggregated data can be made more understandable to the end user, thus enhancing her/his data literacy for computational analysis.

Keywords

linked open data, digital humanities, knowledge graphs, cultural history

1. Introduction

Cultural history data is often stored in separate data sources in different data formats, making collecting and harmonizing data for computational analyzes a tedious process. Transforming digitized archival data into linked data produces machine readable data that is suitable for digital humanities research methods. Such data also follows FAIR Guiding Principles (Findable, Accessible, Interoperable and Re-usable) [1]. Linked archival data also improve knowledge discovery, information search, and retrieval, and helps with data integration and enrichment [2]. In a larger setting, linked datasets form clouds of intelinked knowledge graphs, such as the Linked Open Data (LOD) cloud[1]. In this paper, the SampoSampo cloud of knowledge graphs [3, 4] is considered from a data literacy point of view: how to make aggregated linked data more transparent and understandable to the user. SampoSampo combines alignment data about people, organizations, and places from related cultural history datasets into one knowledge graph, allowing access and computational analysis of these datasets on a global level in new ways. The data service is analogous to VIAF.org [5] that aligns entities of tens of national library collections around the world, but adapted to a dozen instances of the Sampo series of LOD services and semantic portals [6] and a dozen external datasets, such as Wikidata.

For example, letters are often stored in different archives. In the case of LetterSampo Finland [7],

SemDH2026: Third International Workshop of Semantic Digital Humanities
*Corresponding author.
henna.poikkimaki@aalto.fi (H. Poikkimäki); eero.hyvonen@aalto.fi (E. Hyvönen); petri.leskinen@aalto.fi (P. Leskinen)
0000-0003-3362-8438 (H. Poikkimäki); 0000-0003-1695-5840 (E. Hyvönen); 0000-0003-2327-6942 (P. Leskinen)

[1]LOD Cloud: https://lod-cloud.net/

our previous study [8] shows that the letters that an actor A has received are mostly found in one data source (i.e., his/her fond), but by combining multiple letter collections can add the recipients B of A's letters to the actor's personal network using the letters sent by A to B. Such information about how the aggregated data are actually compiled is useful for the end user for a data literacy point of view. Furthermore, it can be difficult to say what type of relationship the actors had based only on letter metadata, and how meaningful the correspondence was to the actors themselves [9]. In that case, it could be of interest to try to look up the actors in other data sources in addition to letter collections to see a possible explanation for the letters sent.

Sampo systems aggregate data from different domains into knowledge graphs. The SampoSampo system acts as an alignment service for entities in different Sampos, and can therefore be of great help in finding connections between entities, such as people, in different domains to enrich the data. In addition to finding explanations for connections found in one data set, one can also find possible connections that are missing in the first data set (e.g., there are no letters between two actors, even though one could assume they are intimately connected based on a different data set). New data sets can also bring new information about the actor, as well as conflicting data (e.g., different dates of birth for an actor). Aggregated data sets offer new possibilities for computational analysis and allow more in-depth explanations of the results.

2. Related Work

Heterogeneous cultural heritage information has been combined and interlinked using semantic web technologies on many occasions, ensuring interoperability of data and improving the unified management, search, and retrieval of cultural heritage data, as well as data integration, data mining, and knowledge extraction [10]. One such system is Europeana which combines cultural heritage content across Europe using linked data principles [11] and Sampo systems listed in table 1.

SampoSampo acts as an alignment service: it collects identifiers of entities in various sources [4], similarly to the Virtual International Authority File (VIAF) and The International Standard Name Identifier (ISNI) [12]. Wikidata is also a relevant addition for authority control [13]. The many roles of Wikidata in digital humanities, ranging from content provider to technology stack, are discussed in [14].

Despite its advantages, Europeana and other digital archives suffer from heterogeneous and inconsistent metadata that has a direct effect on the ability to search, retrieve, and interpret records [15]. Similarly the letter and the writer metadata quality vary in LetterSampo Finland [8], making it difficult to query the letter metadata, e.g., from a group of writers with a certain occupation and hindering temporal analyses of the data as the sending dates of the letters have poor accuracy.

Data literacy, the ability to read, understand, and evaluate data and results of basic analysis as well as use that information for decision-making has become a vitally important skill as the volume of data has increased exponentially [16]. In the context of digital humanities there are several sources of uncertainty: 1) uncertainties within data sources such as missing, ambiguous, conflicting, and unverified data as well as human subjectivity 2) digitization errors 3) datafication issues and 4) statistical models and model appropriateness [17]. It is important to consider these uncertainties when designing analyzes and before making conclusions.

3. SampoSampo Data Service

SampoSampo linked open data service aligns almost 100 000 people, over 55 000 organizations and over 27 000 places from Sampo systems built on data in different domains: academic records, biographies, historical letter collections, art collections, opera performances, parliamentary speeches, military history and Finnish fiction literature. In addition, entities are linked to various other data sources. Data sources and number of people in each source are shown in table 1. The beta version of SampoSampo web portal[2]

[2] https://samposampo.fi/en

is currently available. The data has been published as a linked open data service including as SPARQL endpoint[3], and as data dumps in Zenodo[4].

Each actor in a data source is described as a proxy, and proxies representing the same actor in different data sources are connected to one person type instance, combining URLs of actors from various data sources into one knowledge graph. Systems like SampoSampo make it easier to query information about an actor from different databases, using e.g. federated SPARQL queries [18] if the SPARQL endpoint is available.

SampoSampo also contains almost 200 000 person-place relations derived and deduced from different Sampos, for example actor A has sent letters from place X or actor B has had a career related event in place Y. Similarly, pipelines for obtaining person-person relations from different Sampo systems could be created, and the resulting relations can be used as links between actors in historical social network analysis. In order to study how and with whom the actor interacts in different contexts, queried data could be modeled using multilayer networks where each layer presents interactions of actors in different contexts [19].

Sampo System	People	Other Source	People
LetterSampo Finland [7]	31266	Wikidata.org	48568
AcademySampo [20]	23097	ISNI.org	36093
BiographySampo [21]	21584	KANTO[5]	33031
BookSampo [22]	9841	Wikipedia.org	32920
WarSampo [23]	4154	Geneanet[6]	30195
LetterSampo Finland: J. V. Snellman Letters	3523	VIAF.org	22599
LetterSampo Finland: Albert Edelfelt Letters	3327	Wikitree.com	7377
ParliamentSampo [24]	2111	Geni.com	6059
LetterSampo Finland: Åbo Akademi	1372	ULAN[7]	2292
ArtSampo [25]	1095	HISTO [26]	534
OperaSampo [27]	950		
Norssi High School Alumni [28]	703		

Table 1
Number of people in Sampo systems (first and second columns from the left)) and external data sources (third and fourth columns) from which entities in SampoSampo are found.

4. Overview of the SampoSampo Metadata

Figure 1 shows the actors shared between the data sources. On each row the color represents the proportion of actors in the row data source who appear also in the column data source. The diagonal tells the proportion of actors in the data source who do not appear in any other data sources. Data sources are ordered by the number of actors, and data sources naturally share a larger proportion of actors with larger data sources than smaller ones. If the number of shared actors is relatively high, one might want to consider including both data sets in the analysis for additional information. For example, if one goes through the artist in ArtSampo, in addition to larger data sets, one could check what information is available in ULAN or Albert Edelfelt Letters.

Out of Sampo systems, AcademySampo and WarSampo share the smallest proportion of actors with larger data sources: AcademySampo contains generally people from earlier time period than the big

[3]SampoSampo LOD service: https://www.ldf.fi/dataset/sampo
[4]SampoSampo Zenodo data: https://zenodo.org/records/18188936
[5]https://finto.fi/finaf/en/
[6]https://fi.geneanet.org/
[7]https://www.getty.edu/research/tools/vocabularies/ulan/

data sources (see figure 3), and WarSampo includes soldiers, many of whom may have been general workers rather than politicians or artists and therefore do not appear often in other data sources. Both also have relatively high proportion of unique actors who do not appear in any other data sources. Smaller Letter data sources share most of their person type entities with LetterSampo Finland, as the same letter metadata is also found from there.

Figure 2 shows the number of people, places and organizations of SampoSampo that appear in a certain number of data sources. The majority of people and organizations appear in more than one data source, for people the most common number of data sources is two, and for organizations three. Places usually appear in fewer data sources, but that could change depending on what level the places are considered: in the SampoSampo places vary from very local places like manors to continents. When resources appear in more than one data source, there is a possibility to enrich the information about the resource by combining data sources.

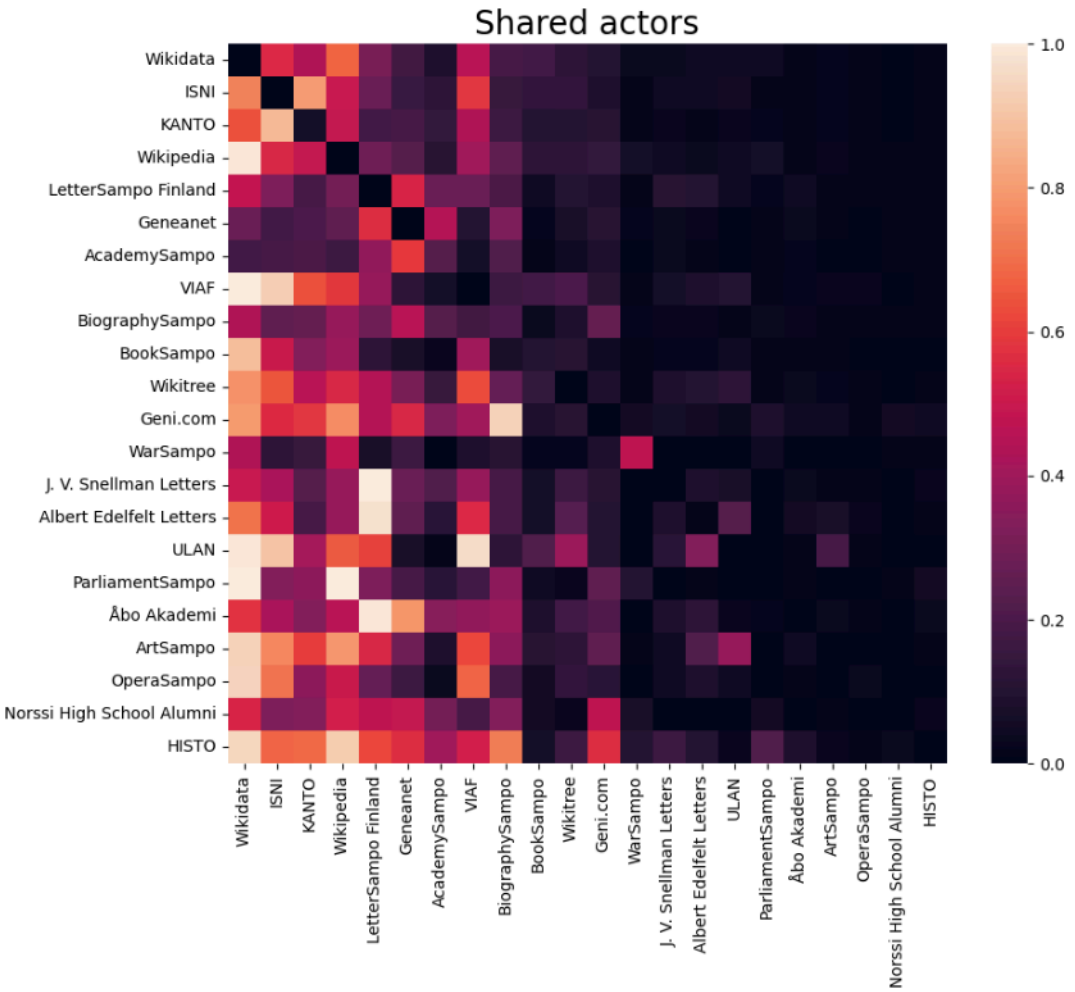

Figure 1: Each cell (i, j) tells the proportion of actors in data source i who appear also in data source j. The diagonal (i, i) tells the proportion of actors in data source i who do not appear in any other data sources.

Because data comes from different domains, the life times of people in different knowledge graphs also varies. One might want to focus on people who were active at the same time: for example, static networks over large observation periods can produce temporally overlapped, difficult to interpret communities as the node pairs are constrained by the corresponding actors' active lives [29]. Figure 3 shows the number of proxies whose time of birth and death is known, who are alive during the years in each data source. For example, AcademySampo and LetterSampo Finland have more people from earlier years, many of whom do not appear in most of the other sources.

As an alignment service, the purpose of SampoSampo is not to bring in all information about the

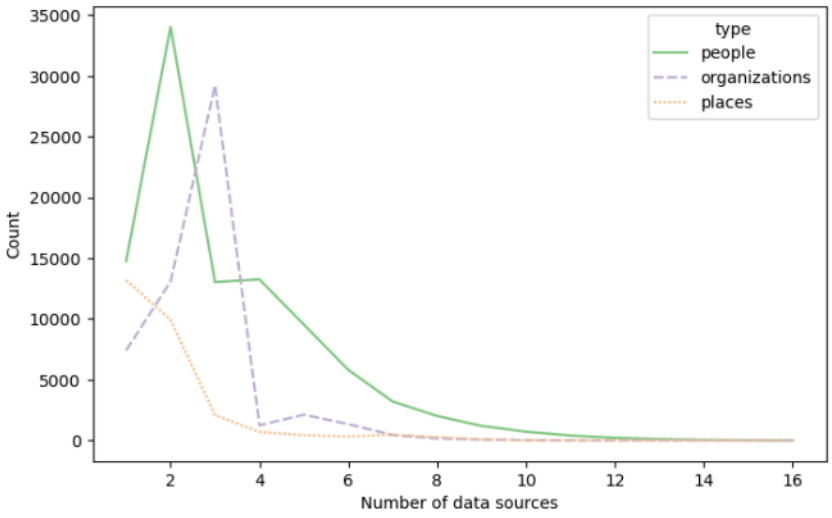

Figure 2: Number of people, places and organizations that appear in a certain number of data sources.

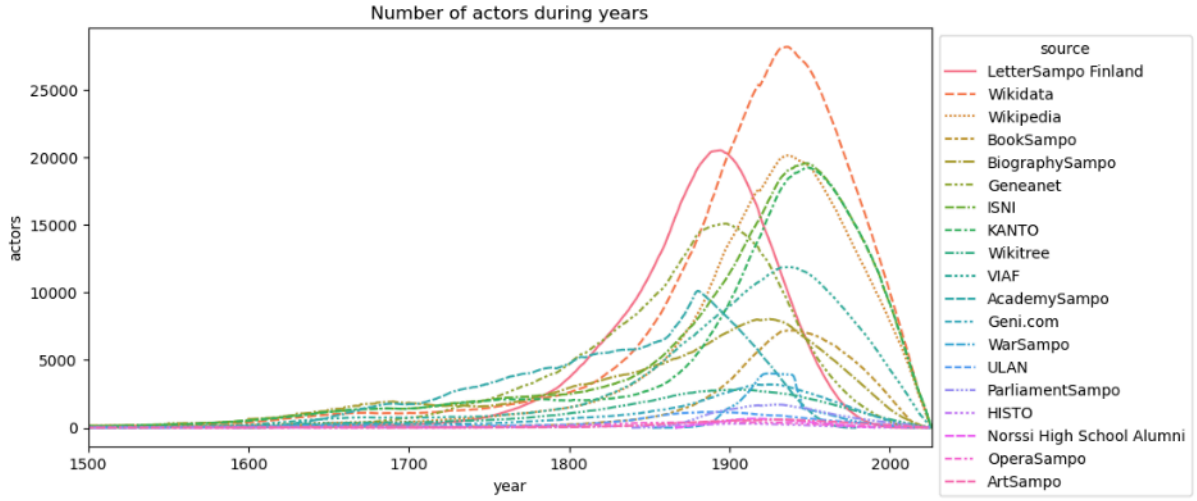

Figure 3: Number of people alive across the years in each data source.

actors, but rather to provide some very basic information that can help with alignments and queries of actors for closer study. However, this information might not be available in all data sources and the information in different data sources might be inconsistent. Figure 4 shows the proportion of actors who have information on time and place of birth and death, sex available, and the number of different types of inconsistencies. The gender of the actor and time of birth and death are well known. The time of birth and death have the most inconsistencies, while gender is rarely conflicting information. The places of birth and death are known for over 60% of actors. In addition, some actors have images available. The process of automatically detecting conflicting information and showing it to end users of data is described in [30]. In addition to different metadata in figure 4, SampoSampo also contains different name variations of actors. To get an idea about the more detailed metadata available and its quality in each data source, one has to consider each data source separately as has been done in the case of LetterSampo Finland [8].

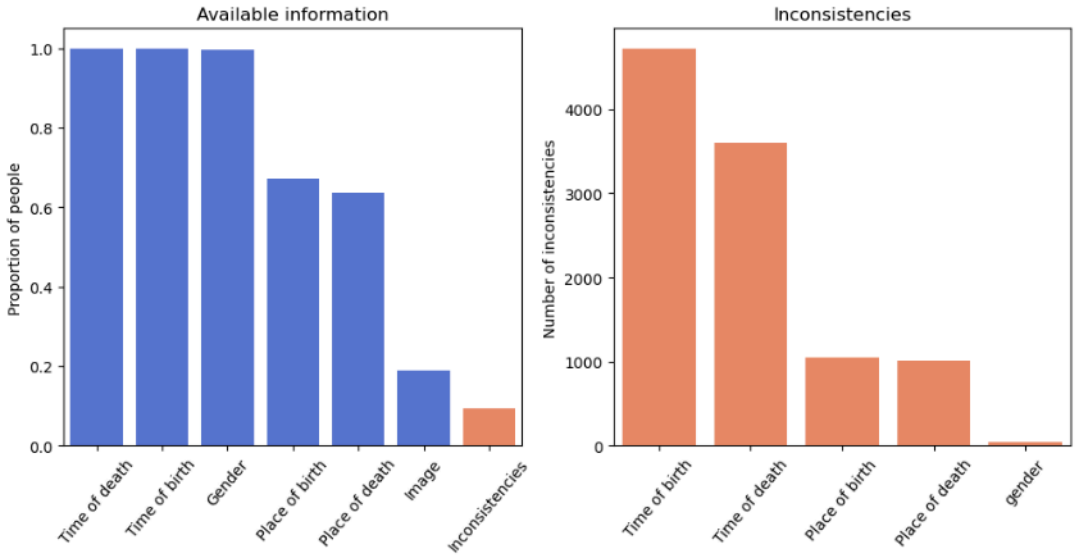

Figure 4: Available information about the actors (beside names and/or name variants) and the number of inconsistencies in the data sources.

5. Discussion and Future Work

This paper discussed the challenge of making data in an interlinked knowledge graph cloud more transparent and understandable to the end user for Digital Humanities research. It was argued that data-analyzes based on a data alignment LOD service can be used for this purpose. The argument was supported by presenting practical examples using the SampoSampo cloud of knowledge graphs and its LOD alignment service.

Poor metadata quality has a negative effect on the ability to query and analyze data from aggregated data sets. It is important to recognize the characteristics and limitations of the data during different stages of the data analysis. SampoSampo combines some metadata about people, places and organizations, and metadata about people was looked into here. Many of the people belong to more than one data source, indicating the sources can enrich each other. Some sources share more actors than the others, which could be a deciding factor on what to include in potential data analyzes. One reason is similar domains that the data sources share, aother reason is the similar time period they cover.

SampoSampo contains a very limited amount of metadata, that could be useful when linking people across the systems and querying them. Times of birth and death are well known, but there are some inconsistencies. Places of birth and death are less known. Gender of the person is in most of the cases known and has rarely inconsistencies across data sources. To further increase understanding about the data in Sampo systems, pipelines and tools for evaluating the metadata quality in different Sampo systems could be developed and also include organizations and places in the quality assessment.

Person-place relations form links between people and places. In future, it would be of interest also to construct pipelines for obtaining person-person relations in different Sampo systems, and use those relations as basis for network analysis. This could share more light with whom some people have interacted with in different contexts, and how combining different Sampo systems enriches personal networks of historical Finnish people.

Acknowledgments

Thanks to Jouni Tuominen, Annastiina Ahola and Heikki Rantala and other co-workers working with SampoSampo. The funding was received from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under decision number 367753 (01.01.2026 – 31.12.2027).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[2] A. Hawkins, Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web, Archival Science 22 (2022) 319–344.

[3] E. Hyvönen, A. Ahola, P. Leskinen, H. Rantala, J. Tuominen, How to create a portal for digital humanities research using a linked open data cloud of cultural heritage knowledge graphs: Case SampoSampo, in: Proceedings of the Second International Workshop of Semantic Digital Humanities (SemDH 2025), co-located with the Extended Semantic Web Conference 2025 (ESWC 2025), volume 4009, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4009/paper_11.pdf.

[4] E. Hyvönen, A. Ahola, P. Leskinen, J. Tuominen, Samposampo: A portal for studying enriched data and semantic connections on a cultural heritage linked open data cloud, in: The Semantic Web: ESWC 2025 Satellite Events, Portoroz, Slovenia, June 1 - 5, 2025, Proceedings, volume 15832 of *Lecture Notes in Computer Science*, Springer-Verlag, 2025, pp. 67–74. URL: https://doi.org/10.1007/978-3-031-99554-5_13.

[5] T. B. Hickey, J. A. Toves, Managing ambiguity in VIAF, DLib Magazine 20 (2014). doi:doi:10.1045/july2014-hickey.

[6] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web 14 (2023) 729–744. doi:10.3233/SW-223034.

[7] E. Hyvönen, P. Leskinen, H. Poikkimäki, H. Rantala, R. Leal, J. Tuominen, S. ˇDrobac, O. Koho, I. Pikkanen, H.-L. Paloposki, Searching, Exploring, and Analyzing Historical Letters and the Underlying Networks: LetterSampo Finland – Finnish 19th-Century Letters on the Semantic Web, Digital Humanities in the Nordic and Baltic Countries Publications 7 (2026). URL: https://journals.uio.no/dhnbpub/article/view/12932. doi:10.5617/dhnbpub.12932.

[8] H. Poikkimäki, P. Leskinen, E. Hyvönen, "Analyzing Aggregated Knowledge Graphs on a Global Level for Better Data Literacy: Case LetterSampo Finland", in: B. Villazón-Terrazas, F. Ortiz-Rodriguez, S. Tiwari, T. Riechert, E. Marx (Eds.), Knowledge Graphs and Semantic Web, Springer Nature Switzerland, Cham, 2026, pp. 265–279.

[9] R. M. Morrissey, Archives of connection, Historical Methods: A Journal of Quantitative and Interdisciplinary History 48 (2015) 67–79. URL: https://doi.org/10.1080/01615440.2014.962208. doi:10.1080/01615440.2014.962208.

[10] E. Moraitou, J. Aliprantis, Y. Christodoulou, A. Teneketzis, G. Caridakis, Semantic bridging of cultural heritage disciplines and tasks, Heritage 2 (2019) 611–630. URL: https://www.mdpi.com/2571-9408/2/1/40. doi:10.3390/heritage2010040.

[11] A. L. Silva, A. L. Terra, Cultural heritage on the semantic web: The europeana data model, IFLA Journal 50 (2024) 93–107. URL: https://doi.org/10.1177/03400352231202506. doi:10.1177/03400352231202506.

[12] A. Angjeli, A. Mac Ewan, V. Boulet, Isni and viaf–transforming ways of trustfully consolidating identities, IFLA WLIC 2 (2014).

[13] C. Bianchini, S. Bargioni, C. C. P. di San Girolamo, Beyond viaf: Wikidata as a complementary tool for authority control in libraries, Information Technology and Libraries 40 (2021).

[14] F. Zhao, A systematic review of wikidata in digital humanities projects, Digital Scholarship in the Humanities 38 (2022) 852–874. URL: https://doi.org/10.1093/llc/fqac083. doi:10.1093/llc/fqac083. arXiv:https://academic.oup.com/dsh/article-pdf/38/2/852/50488385/fqac083.pdf.

[15] C. Papaioannou, Historical research in the digital age: Opportunities, challenges, and critical reflections through the case of europeana, Journal of Integrated Information Management 10 (2025) 19–24. URL: https://ejournals.epublishing.ekt.gr/index.php/jiim/article/view/41305. doi:10.26265/jiim.v10i1.41305.

[16] Y. Cui, F. Chen, A. Lutsyk, J. P. Leighton, M. Cutumisu, Data literacy assessments: a systematic literature review, Assessment in Education: Principles, Policy & Practice 30 (2023) 76–96. URL: https://doi.org/10.1080/0969594X.2023.2182737. doi:10.1080/0969594X.2023.2182737.

[17] G. Panagiotidou, H. Lamqaddam, J. Poblome, K. Brosens, K. Verbert, A. Vande Moere, Communicating uncertainty in digital humanities visualization research, IEEE Transactions on Visualization and Computer Graphics 29 (2023) 635–645. doi:10.1109/TVCG.2022.3209436.

[18] M. Acosta, O. Hartig, J. Sequeda, Federated RDF Query Processing, Springer International Publishing, Cham, 2019, pp. 754–761. URL: https://doi.org/10.1007/978-3-319-77525-8_228. doi:10.1007/978-3-319-77525-8_228.

[19] M. E. Dickison, M. Magnani, L. Rossi, Multilayer social networks, Cambridge University Press, 2016.

[20] P. Leskinen, E. Hyvönen, Linked open data service about historical finnish academic people in 1640–1899, Digital Humanities in the Nordic and Baltic Countries Publications 3 (2020) 284–292. URL: https://journals.uio.no/dhnbpub/article/view/11199. doi:10.5617/dhnbpub.11199.

[21] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, Biographysampo - publishing and enriching biographies on the semantic web for digital humanities research, in: P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, K. Hammar (Eds.), The Semantic Web. ESWC 2019, Springer-Verlag, 2019, pp. 574–589. URL: https://doi.org/10.1007/978-3-030-21348-0_37. doi:10.1007/978-3-030-21348-0_37.

[22] E. Hyvönen, A. Ahola, E. Ikkala, Booksampo fiction literature knowledge graph revisited: Building a faceted search interface with seamlessly integrated data-analytic tools, in: Theory and Practice of Digital Libraries (TPDL 2022), Accelerating Innovations Track, Padova, Italy, Springer, 2022, p. 506–511. URL: https://doi.org/10.1007/978-3-031-16802-4_54.

[23] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, Warsampo knowledge graph: Finland in the second world war as linked open data, Semantic Web – Interoperability, Usability, Applicability 12 (2021) 265–278. URL: https://doi.org/10.3233/SW-200392. doi:10.3233/SW-200392.

[24] E. Hyvönen, L. Sinikallio, P. Leskinen, S. Drobac, R. Leal, M. L. Mela, J. Tuominen, H. Poikkimäki, H. Rantala, Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland, Semantic Web 16 (2025) SW–243683. URL: https://journals.sagepub.com/doi/abs/10.3233/SW-243683. doi:10.3233/SW-243683.

[25] A. Ahola, H. Rantala, E. Hyvönen, Artsampo - finnish art on the semantic web, in: The Semantic Web: ESWC 2024 Satellite Events, Hersonissos, Crete, Greece, May 26 - 30, 2024, Proceedings, Springer, 2024.

[26] H. Rantala, E. Hyvönen, E. Ikkala, Creating the histo ontology of finnish history events, in: Data for History 2021: Modelling Time, Places, Agents, 2021. URL: https://d4h2020.sciencesconf.org/data/pages/Rantala_Hyvo_nen_Ikkala_HISTO_Ontology_2.pdf, abstract.

[27] A. Ahola, E. Hyvönen, H. Rantala, A. Kauppala, Historical opera and music theatre performances on the semantic web: Operasampo 1830-1960, in: Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI. Proceedings of the 20th International Conference on Semantic Systems, 17-19 September 2024, Amsterdam, The Netherlands, IOS Press, 2024, pp. 386–402. URL: https://doi.org/10.3233/SSW240031, dOI: 10.3233/SSW240031.

[28] E. Hyvönen, P. Leskinen, E. Heino, J. Tuominen, L. Sirola, Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the semantic web, in: Proceedings, Language, Data and Knowledge (LDK 2017), Springer-Verlag, 2017, pp. 113–119. URL: https://doi.org/10.1007/978-3-319-59888-8_9.

[29] J. Ureña-Carrion, P. Leskinen, J. Tuominen, C. van den Heuvel, E. Hyvönen, M. Kivelä, Communication now and then: Analyzing the republic of letters as a communication network, Applied Network Science 7 (2022). URL: https://doi.org/10.1007/s41109-022-00463-1.

[30] Petri Leskinen and Annastiiina Ahola and Heikki Rantala and Jouni Tuominen and Eero Hyvönen, Consistency checking in a cloud of interlinked cultural heritage knowledge graphs – first results of using the samposampo data service and portal, in: Proceedings of DHNB-2026, March 9-13, 2026, Aarhus, Denmark, 2026. In press.