Digital Humanities on the Semantic Web: from Infrastructure to Practical Applications, AI-based Knowledge Discovery, and Web of Wisdom

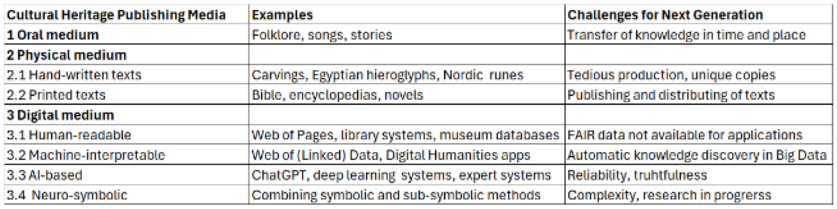
Eero Hyvönen[0000–0003–1695–5840]

Aalto University, Department of Computer Science and
University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)
eero.hyvonen@aalto.fi https://seco.cs.aalto.fi/eahyvone/

Abstract. This extended abstract of a keynote talk at the 29th International Conference on Theory and Practice of Digital Libraries (TPDL 2025) presents the vision, research, and lessons learned in Finland 2003–2025 on building a national level Semantic Web infrastructure for publishing Cultural Heritage (CH) content and Digital Humanities research. Lessons learned are presented on how the infra has been used to create over 20 online applications that have been visited by millions of users. By using AI-based methods, steps towards a Web of Wisdom can be taken. The notion of explainable relational search is considered as an example. The work presented is unique due to its systematic national level nature and long time span of over twenty years.

Keywords: knowledge graphs · relational search · knowledge discovery · information retrieval · large language models · generative AI

1 Media Generations for Publishing CH Content



Cultural Heritage Publishing Media	Examples	Challenges for Next Generation
1 Oral medium	Folklore, songs, stories	Transfer of knowledge in time and place
2 Physical medium		
2.1 Hand-written texts	Carvings, Egyptian hieroglyphs, Nordic runes	Tedious production, unique copies
2.2 Printed texts	Bible, encyclopedias, novels	Publishing and distributing of texts
3 Digital medium		
3.1 Human-readable	Web of Pages, library systems, museum databases	FAIR data not available for applications
3.2 Machine-interpretable	Web of (Linked) Data, Digital Humanities apps	Automatic knowledge discovery in Big Data
3.3 AI-based	ChatGPT, deep learning systems, expert systems	Reliability, truhtfulness
3.4 Neuro-symbolic	Combining symbolic and sub-symbolic methods	Complexity, research in progrerss

Fig. 1. Generations of media for transferring knowledge and Cultural Heritage content

Cultural Heritage knowledge has been transferred between people and their generations over time using various media, as illustrated in Fig. 1. First, knowledge was transferred orally, but this was challenging as speech could reach only nearby people and it was not possible to store content outside human brains.

To solve the issue, writing systems using physical media such as clay tables and papyrus sheets were developed, but using them required a lot of manual work. As a remedy, book printing was invented in China and Europe, but the use of physical media was not fast and effective enough in the globalizing world. To address the issue, digitalization and using the Web to publish and distribute content for humans to read and see was the next answer. However, as computers were used more and more, the need for publishing content for not only humans but also for data analyses and applications to use crew, which led to the era of publishing FAIR data and Web of Data. When more and more Big Data became available, it became more and more difficult to use it. Automatic methods were needed to help the user leading to the bloom of AI-based systems, first in the 1980's to knowledge-based systems using symbolic methods, and today especially systems based of sub-symbolic neural computing and deep learning. A major technological challenge to be addressed today is how to mitigate knowledge acquisition challenges of symbolic and hallucination issues of sub-symbolic AI methods in the era of neuro-symbolic hybrid systems.

This paper reviews work and lessons learned in Finland pertaining to the transition from a human-readable Web to a machine-interpretable intelligent Semantic Web, and towards a Web of Wisdom based on AI and neuro-symbolic computing. The application domain is Cultural Heritage data and Digital Humanities research. It is argued that a national Semantic Web infrastructure based on and extending the international W3C and other standards are needed for the purpose, including ontology and data services, and tools for easy application development. Results on developing a series of in-use applications on top of the infrastructure are presented.

2 A Semantic Web Infrastructure for Digital Humanities

Ontology Services The Semantic Web (SW) sees the Web as an interlinked collection of data, Web of Data, in addition to the traditional space of interlinked hypertext documents, Web of Pages. Its technological basis are recommendations of the W3C, i.e., the "layer cake model", that lays out a basis of shared semantics for interoperability of data[1]. Founded on first order predicate logic, the semantics of the SW are independent of application domains and natural languages used on the Web. However, to develop applications domain and application specific ontologies and data are needed, too [3]. This involves 1) metadata models for representing content, 2) vocabularies (knowledge organization systems) for populating the metadata models, and 3) Linked Open Data (LOD) to be used and re-used in applications. In order to serve and (re-)use the data resources 1–3 efficiently, public ontology services and and data services are needed as well as various tools for creating and linking data and building the applications.

Building a SW infra on a national level started in Finland as the FinnONTO series of projects[2] 2003–2012 that created the prototype of a public ontology

[1] W3C Semantic Web recommendations: https://www.w3.org/2001/sw/wiki/Main_Page

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

service ONKI.fi with a cloud of interlinked ontologies, based on in-use domain-specific thesauri. ONKI.fi was deployed in 2014 by the National Libary as the current in-use Finto.fi service.

Linked Open Data Services Research in FinnONTO continued to developing a similar kind of service for data and metadata schema publishing: The Linked Data Finland platform LDF.fi. Based on its SPARQL endpoint and other services, a series of over 20 semantic "Sampo" portals for Digital Humanities have been published on the Web that have had millions of users in total. A lesson learned in this work is that key challenges of re-using LOD include missing data schemas, formal quality of data, and truthfulness of the data. To encourage data publishers to address these issues, an 8-start model was proposed in LDF.fi [6] extending the 5-star model[3] coined by Tim Berners-Lee: **6th star**: Describe explicitly and publish the schemas used in the datasets. **7th star**: Explain the formal quality of the dataset w.r.t. the schemas used, so that the user can tell when the data quality matches her needs. **8th star**: Give explanations when the data is factually correct with respect to the real world and when not.

Sampo Model and Sampo-UI Framework More and more LOD were published via SPARQL endpoints by re-using the infrastructure above. At the same time, the challenge of creating User Interfaces (UI) easily for searching, browsing, and analyzing the LOD became ever more important. Our solution approach to address the UI challenge is the Sampo-UI framework [7, 8]. Its key features include: 1) Sampo-UI makes it possible to create faceted semantic search-based applications integrated seamlessly with data-analytics visualization tools needed, e.g., in Digital Humanities research. 2) Based on the Sampo model [2], the *data service is completely separated* from the UI design via the SPARQL API only. This makes it possible to create UIs for external endpoints. 3) The UI is created in a *declarative fashion* by modifying a set of JSON configuration files and SPARQL queries [8] that adapt the UI to the underlying data model and ontologies. 4) To facilitate extremely quick prototyping, an *existing "vanilla" Sampo-UI application is used as a starting point* and adapted to the new application by modifying the configuration files and SPARQL queries.

The Sampo portals[4] are overviewed in [2] and the SW infrastructure in [3].

3 Towards Explainable Knowledge Discovery

Digital media (category 3 in Fig. 1) is based on data. According to the Data-Information-Knowledge-Wisdom (DIKW) hierarchy of data science [9], new value is created as 'data' (know just the data, nothing about it) changes into 'information' (know what the data is), then into 'knowledge' (know how the information is used), and finally into 'wisdom' (know why; explaining knowledge) (cf. Fig. 2). This transition is happening on the Web that is gradually changing from a data/information publishing platform into a knowledge base, and finally into an intelligent question answering system, a Web of Wisdom (WoW) [4].

[3] https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/

[4] Sampo portals on the Web: https://seco.cs.aalto.fi/applications/sampo/

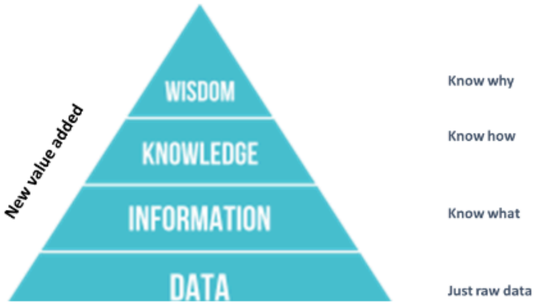


Fig. 2. Data-Information-Knowledge-Wisdom (DIKW) hierarchy of data science

The Sampo portals use machine-interpretable data as their medium (3.2 in Fig. 1) and can also be considered simple AI-based systems (3.3), because they do simple logic reasoning using the SW semantics and enrich data by linking. AI-based methods were also used for enriching primary textual data by Named Entity Recognition and Linking, automatic keyword extraction, and topic detection. Furthermore, the data-analytic tools and visualizations available in the portals help the user in finding interesting patterns of knowledge in the data. However, the tools are used and results interpreted by humans: they are not automatic agents with the wisdom capable of explaining their knowledge.

A way to make a step towards the WoW on the Semantic Web is *relational search* (RS) [4] where the notion of search is extended to finding "interesting" or even serendipitous connections between the resources in a knowledge graph (KG), such as persons, places, and events—with explanations. For example: "How are German novelists of the 19th century related to France?" Such semantic connections can be based on various criteria: German people (or their family members) were born or died in Paris, French topics were discussed in their novels, they wrote a novel or an article in French, their publisher was a French company, their portraits are in Louvre, they got a medal of honor in Lyon, etc.

There are different approaches to relational knowledge discovery. *Domain agnostic RS* is based on generic graph algorithms re-usable in different application domains while in *knowledge-based RS* symbolic knowledge representations and reasoning are used to capture the semantics "interestingness" or "serendipity" more precisely and to explain the relations in natural language. A challenge of knowledge-based methods is in knowledge acquisition needed for formulating the rules. An approach to address this problem is to use available linguistic texts and machine learning as a basis for providing explanations. This leads to *linguistics-based RS* approaches to solve relational search problems. This approach of building explanations from text corpora bears similarity with the idea of using LLMs for question answering. It is possible to solve relational search problems by simply asking ChatGPT-like systems to explain relationships by asking questions, such as "How is Adolf von Becker related to Switzerland?".

A new "meta-sampo" *SampoSampo – Connecting Everything to Everything Else* is underway on top of the Sampo systems on the Web [5]. A novelty of SampoSampo is to use knowledge-based and linguistic relational search for knowledge discovery on top of the cloud of interlinked Sampo KGs and LOD data services.

4 Conclusions

The Web has become a major media for Cultural Heritage knowledge transfer that is independent of time and place, changing gradually from a Web of Pages into a Web of Wisdom. Here the information needs of end users are addressed not only by retrieving documents or analyzing data, but also directly by intelligent problem solving and question answering, based on a merger of symbolic and sub-symbolic methods of AI. This is a new possibility that the Semantic Web, "a new form of Web content that is meaningful to computers" was predicted to unleash in 2001, boosted today by recent developments in LLMs based on deep learning and textual data. As an example, the idea of using relational search in knowledge discovery was considered in this paper. Relational search can use complementary symbolic and sub-symbolic methods: domain specific knowledge and KGs can be to used to find truthful relations, answer quantitative questions, create textual answers, and use data analyses and visualizations for studying the relations, while LLM-based methods are able to use large collections of textual data for question answering overcoming the knowledge acquisition bottleneck of knowledge-based systems, but with unpredictable hallucinations.

A possibility for enhancing the capabilities of LLMs would be to finetune them for particular domain areas with additional contextual information using domain specific training materials. Such training material could be generated from new texts but also from structured data sources, such as KGs. However, finetuning the very large LLMs is challenging in many ways, and using prompt engineering[5] techniques such as Retrieval-Augmented Generation (RAG) would provide an easier way to get better answers. Here prompting can be enhanced by using contextual data from KGs. There is also the possibility to apply LLMs and Generative AI to creating richer structured data in KGs [1].

Our work to develop the Finnish LOD infrastructure and its applications continues as part of the Finnish FIN-CLARIAH program[6] in relation to the Pan-European CLARIN and DARIAH infrastructures (CLARIAH=CLARin+darIAH).

Acknowledgments. The work presented is based on research conducted at the Semantic Computing Research Group (SeCo)[7] at the Aalto University and University of Helsinki, Finland. Our research infrastructure work is funded today by the Research Council of Finland as part of the NextGenerationEU program. The author thanks the Finnish Cultural Foundation for an Eminentia Grant for reflecting the line of research discussed.

[5] Prompt engineering guide: https://www.promptingguide.ai/

[6] LOD part of FIN-CLARIAH: https://seco.cs.aalto.fi/projects/fin-clariah/

[7] SeCo Research Group homepages: https://seco.cs.aalto.fi/

Bibliography

[1] Ahola, A., Peura, L., Leal, R., Rantala, H., Hyvönen, E.: Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in digital humanities. In: Proc. of the 2nd Int. Conf. on Data & Digital Humanities Generative AI for Text and Multimodal Data, Portugal (2024), https://seco.cs.aalto.fi/publications/2024/ahola-et-al-genai-2024.pdf, forth-coming

[2] Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web **14**(4), 729–744 (2022). https://doi.org/10.3233/SW-223034

[3] Hyvönen, E.: How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web. Semantic Web **15**(4) (2024). https://doi.org/10.3233/SW-243468

[4] Hyvönen, E.: Serendipitous knowledge discovery on the web of wisdom based on searching and explaining interesting relations in knowledge graphs. Journal of Web Semantics (2025), https://doi.org/10.1016/j.websem.2024.100852

[5] Hyvönen, E., Ahola, A., Leskinen, P., Rantala, H., Tuominen, J.: How to create a portal for digital humanities research using a linked open data cloud of cultural heritage knowledge graphs: Case SampoSampo. In: Proceedings: SemDH 2025 Second International Workshop of Semantic Digital Humanities, Portoroz, Slovenia. CEUR-WS.org (2025), https://seco.cs.aalto.fi/publications/2025/hyvonen-et-al-samposampo-semdh-2025.pdf, in press

[6] Hyvönen, E., Tuominen, J.: 8-star linked open data model: Extending the 5-star model for better reuse, quality, and trust of data. In: Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024). vol. 3759. CEUR-WS.org (September 2024), https://ceur-ws.org/Vol-3759/paper4.pdf

[7] Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web **13**(1), 69–84 (2022). https://doi.org/10.3233/SW-210428

[8] Rantala, H., Ahola, A., Ikkala, E., Hyvönen, E.: How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In: Proc. of 8th Int. Workshop on the Visualization and Interaction for Ontologies and Linked Data, Athens, Greece. vol. 3508. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3508/paper3.pdf

[9] Rowley, J.E.: The wisdom hierarchy: representations of the DIKW hierarchy. Journal of Information Science **33**, 163—180 (2007), https://api.semanticscholar.org/CorpusID:17000089