

8-star Linked Open Data Model: Extending the 5-star Model for Better Reuse, Quality, and Trust of Data

Eero Hyvönen^{1,2,*}, Jouni Tuominen^{3,2,1}

¹*Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*

²*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

³*Helsinki Institute for Social Sciences and Humanities (HSSH), University of Helsinki, Finland*

Abstract

This paper argues, based on lessons learned in developing several in-use Linked Open Data (LOD) services and applications, that biggest challenges of re-using LOD are missing data schemas, formal quality of data, and trust on the correctness of data with respect to the real world. To encourage data publishers to address the issues, three more stars are proposed for the classical 5-star model coined by Tim Berners-Lee. The proposed model is supported by the Linked Data Finland platform, a living lab environment in use supporting LOD publication for data-driven research and application development.

Keywords

Linked open data, Semantic web, 5-star open data, Publishing model

1. Linked Data Principles and 5-star Model

Linked Data (LD) is based on the standards of the W3C, such as RDF(S), SKOS, and OWL. In addition, there are more informal best practices whose intention is to guide data publishers and end users of LD [1]. Most notably, the four linked data principles¹ state that 1) one should use URIs as names for things, 2) use HTTP URIs so that people can look up those names, 3) When someone looks up a URI, one should provide useful information, and 4) include links to other URIs so that they can discover more things. Based on these principles, the 5-star model² was devised in order to encourage data publishers to provide their data in a maximally useful form and way. This idea is analogous to the 5-star rating system in use for hotels; one gets more stars by providing more and better services:

- ★ Make your stuff available on the Web (whatever format) under an open license.
- ★★ Make it available as structured data (e.g., Excel instead of image scan of a table).
- ★★★ Make it available in a non-proprietary open format (e.g., CSV instead of Excel).
- ★★★★ Use URIs to denote things, so that people can point at your stuff.
- ★★★★★ Link your data to other data to provide context.


SEMANTiCS 2024, Posters and Demos

*Corresponding author.

✉ eero.hyvonen@aalto.fi (E. Hyvönen); jouni.tuominen@helsinki.fi (J. Tuominen)

🌐 <https://seco.cs.aalto.fi/u/eahyvone> (E. Hyvönen); <https://seco.cs.aalto.fi/u/jwtuomin> (J. Tuominen)

🆔 0000-0003-1695-5840 (E. Hyvönen); 0000-0003-4789-5676 (J. Tuominen)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Linked Data Principles: <https://www.w3.org/wiki/LinkedData>

²5-star model: <https://5stardata.info/en/>

Three first stars encourage opening data and the two last ones on using LOD technologies.

A most important functionality of LOD services is the provision of a SPARQL endpoint. There are lots of triplestore frameworks available for this³ and SPARQL endpoints on the Web⁴. LOD services typically provide also other services, such as: 1) Dereferencing services of URIs to access the machine/human-readable RDF/HTML data. 2) RDF browsing services in HTML in order to investigate the data. 3) Services for uploading and downloading data. 4) Data management services, for, e.g., editing the data in native form.

2. 7-star model for LOD Publishing

A key challenge in many LOD services is that they provide little information about the data schemas (data models, ontologies) used, which would be important for external re-use. The schemas should also comply with the data provided, which can be a challenge even with simple schemas such as SKOS [2]. To encourage data publishers to address the issues of missing schemas and low data quality w.r.t. to them, we proposed adding two extra stars in the 5-star model [3]. There are, after all, arguable also 7-star hotels around⁵.

★★★★★ Describe explicitly and publish the schemas used in the dataset (unless the schemas are already available elsewhere).

★★★★★ Explain the quality of the dataset against the schemas used in it, so that the user can tell whether the data quality matches her needs.

Earning the 6th star is fairly easy if one has the schema available in a machine-readable form. It is also possible to address the schema issue by using automatic documentations systems, such as LODE⁶ [4]. Getting the 7th star is a bit more challenging. However, two frameworks and standards for validating LD formally based on schema templates (constraints) have emerged: the Shapes Constraint Language (SHACL)⁷, recommendation (i.e., standard) of the W3C, and Shape Expressions (ShEx)⁸ [5]. A benefit of ShEx for our purpose is that its schema templates document in a fairly straightforward way metadata models, so the 6th star can be obtained easily. Furthermore, there are ShEx validators that can be used for obtaining the 7th star in a formal way. A benefit of SHACL is that its constraint language is more versatile and can be used, e.g., for stating and validating constraints between metadata element values.

3. 8-star Model for LOD Publishing

Even for 7-star LOD an important issue hindering the reuse of data remains: can the data be trusted, i.e., does it conform to the facts of the real world it is supposed to model? We argue that it makes sense to separate this dimension of data quality from validating the data formally

³W3C list of large triplestores: <https://www.w3.org/wiki/LargeTripleStores>

⁴W3C list of some SPARQL endpoints: <https://www.w3.org/wiki/SparqlEndpoints>

⁵Such as Burj Al Arab In Dubai and others, cf. <https://traveltriangle.com/blog/7-star-hotels-in-the-world/>

⁶LODE homepage: <https://essepuntato.it/lode/>

⁷SHACL: <https://www.w3.org/TR/shacl/>

⁸ShEx: <https://github.com/shexSpec/shex/wiki/ShEx>

against the schemas. In our work on creating LOD services and semantic portals [6], we have encountered different situations where the data is not trustworthy. For example, a central dataset within the WarSampo knowledge graph [7] is a database of all fallen Finnish soldiers during the Second World War. This data conformed to the metadata model of death records obtained from the data owner National Archives of Finland. To test validity of the data against constraints of the real world, some domain specific SPARQL queries were designed to check potential anomalies, such as if a person participated in an event after his/her death. It turned out that there were soldiers, according to the data, who were wounded after they were killed. Such errors may arise from several reasons. The error may be due to primary data that in the case of war data may be uncertain in many ways, or be due to typing errors when cataloging the data. Another small-scale experiment was made using the Portable Antiquities Scheme (PAS) dataset of the British Museum in PASampo [8]. The PAS uses a broad period classification for archaeological finds, following the FISH (Forum on Information Standards in Heritage) vocabularies, and also includes (chronological) date ranges for the finds ('from date' – 'to date'). The latter allows for more precise temporal recording than using the broad period classification. A SHACL shape was created, that states that the 'from date' has to be 'less than or equal to' the 'to date'. When validating the data with a SHACL validator, some 2000 errors were found, such as an Iron Age quarter stater recorded with the date range AD 80–60 while the correct dating is 80–60 BC.

A source of errors in data is mistakes made when annotating data automatically using, e.g., Named Entity Recognition (NER) and Linking (NEL) methods [9]. For example, in BiographySampo [10, 11] named entities are recognized and linked from textual biographies as new metadata. The same challenge arises when using automatic classification systems for extracting additional metadata, e.g., topical categories for records like for plenary speeches of parliamentarians in ParliamentSampo [12]. In this kind of cases it is often impossible to check correctness of NER/NEL. However, one could at least explicitly state what part of the data has been created automatically and is therefore not necessarily trustworthy.

The notion of truth can be vague and there may be several truths around depending on the point of view. For example, when developing the WW1 application regarding atrocities and casualties of the Great War [13], the numbers given in British and German primary documents were different. This could be due to propaganda, but not necessarily, because the data has been acquired from different sides of the front line. When dealing with historical data, situations are common where nobody really may know what is true, and there are only opinions available. This challenge is also encountered in contemporary data and ontologies. For example, whether Taiwan is considered as part of China or not depends on a political interpretation.

A good question is how can trustworthiness of data be guaranteed and represented so that re-users of data can decide whether to trust the data, a goal of the W3C Credible Web Community Group⁹. We are living in the era of fake news, misinformation and disinformation. Data may be published based on conspiracy theories that cannot be trusted or by people who intentionally publish false data for their own benefit. In the Semantic Web layer cake models [14], trust is on the highest level above (first order predicate) logic that forms the semantic basis of the Semantic

⁹W3C Credible Web Community Group: <https://www.w3.org/community/credibility/>

Web. For example, the arguments of the Flat Earth Society¹⁰ may be logical but the foundational facts are wrong and the data cannot be trusted. Research about trust on the Semantic Web has focused first on digital signatures, certificates, and authentication [15], but later also on detecting and representing factual correctness (veracity) of the data [16].

In spite of the semantic challenges of the notion of truth, one should strive to create data and ontologies that match the known facts of the real world, even if trustworthiness cannot be fully proven. To encourage data publishers towards this the 8th star is proposed in our model:

★★★★★★ Give explanations when the data is factually correct with respect to the real world and when possibly not.

The 8-star model is supported as part of the Linked Data Finland platform LDF.fi¹¹ [3] that allows publication of schemas alongside the actual data as well as automatic data documentation¹² for the 6th star. LDF.fi has been used for publishing lots of datasets as part of a national Finnish LOD infrastructure [17], supporting at the same time also over 20 in-use semantic Sampo portals for Cultural Heritage data and Digital Humanities research [6]. In LDF.fi each dataset is described using VoID¹³, where a rating of 1–8 stars can be given, too. Based on this metadata a homepage for the dataset with a SPARQL endpoint, associated LOD services, and instructions for re-using the data are automatically created [3].

4. Conclusions and Acknowledgements

This paper identified three major bottlenecks for re-using LOD, a major goal of the FAIR principles: missing data schema information, formal quality of data, and trustworthiness of LOD. To encourage data publishers to address these issues, an extension to the classical 5-star model was proposed, and a platform supporting the 8-star model in use was shortly described.

Our work was partially funded by the European Union – NextGenerationEU instrument (grant P3C3I6 by the Research Council of Finland) and the Finnish Cultural Foundation. CSC – IT Center for Science has provided computational resources for our work.

References

- [1] T. Heath, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space* (1st edition), Morgan & Claypool, 2011. URL: <http://linkeddatabook.com/>.
- [2] O. Suominen, E. Hyvönen, Improving the quality of SKOS vocabularies with Skosify, in: *Knowledge Engineering and Knowledge Management. EKAW2012*, Springer, 2012, pp. 383–397. doi:10.1007/978-3-642-33876-2_34.
- [3] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, *Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets*, in: *The Semantic Web: ESWC 2014 Satellite Events*, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.

¹⁰Flat Earth Society homepage: <https://theflatearthsociety.org>

¹¹Linked Data Finland platform: <https://ldf.fi>

¹²Using pyLODE tool <https://github.com/RDFLib/pyLODE> based on LODE

¹³VoID Vocabulary: <https://www.w3.org/TR/void/>

- [4] S. Peroni, D. Shotton, F. Vitali, The Live OWL Documentation Environment: A tool for the automatic generation of ontology documentation, in: *Knowledge Engineering and Knowledge Management. EKAW2012*, Springer, 2012, pp. 398–412. doi:10.1007/978-3-642-33876-2_35.
- [5] J. E. Labra Gayo, E. Prud'hommeaux, I. Boneva, D. Kontokostas, Validating RDF Data, volume 7 of *Synthesis Lectures on the Semantic Web: Theory and Technology*, Morgan & Claypool, 2017. doi:10.2200/s00786ed1v01y201707wbe016.
- [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] 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* 12 (2021) 265–278. doi:10.3233/SW-200392.
- [8] M. Lewis, E. Oksanen, F. Ehrnsten, H. Rantala, J. Tuominen, E. Hyvönen, The impact of human decision-making on the research value of archaeological data, 2024. URL: <https://seco.cs.aalto.fi/publications/2024/lewis-et-al-pasampo-2024.pdf>, in review.
- [9] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: A survey, *Semantic Web* 11 (2020) 255–335. doi:10.3233/SW-180333.
- [10] 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: *The Semantic Web. ESWC 2019*, Springer, 2019, pp. 574–589. doi:10.1007/978-3-030-21348-0_37.
- [11] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case national biography of Finland, *Semantic Web* 14 (2023) 385–419. doi:10.3233/SW-222887.
- [12] 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* (2024). In open review: <https://www.semantic-web-journal.net/content/publishing-and-using-parliamentary-linked-data-semantic-web-parliamentsampo-system>.
- [13] E. Mäkelä, J. Törnroos, T. Lindquist, E. Hyvönen, WW1LOD: An application of CIDOC-CRM to World War 1 linked data, *International Journal on Digital Libraries* 18 (2017) 333–343. doi:10.1007/s00799-016-0186-2.
- [14] R. Goebel, S. Zilles, C. Ringlstetter, A. R. Dengel, G. A. Grimnes, What is the role of the semantic layer cake for guiding the use of knowledge representation and machine learning in the development of the semantic web?, in: *AAAI Spring Symposium: Symbiotic Relationships between Semantic Web and Knowledge Engineering*, 2008, pp. 45–50.
- [15] J. Golbeck, B. Parsia, J. Hendler, Trust networks on the semantic web, in: *Cooperative Information Agents VII*, Springer, 2003, pp. 238–249. doi:10.1007/978-3-540-45217-1_18.
- [16] R. Denaux, J. M. Gomez-Perez, Linked credibility reviews for explainable misinformation detection, in: *The Semantic Web – ISWC 2020*, Springer, 2020, pp. 147–163. doi:10.1007/978-3-030-62419-4_9.
- [17] E. Hyvönen, How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web, *Semantic Web* (2024). doi:10.3233/SW-243468, accepted.