Semantic Web 0 (2014) 1–0
IOS Press

Five Stars of Linked Data Vocabulary Use

Editorial

Krzysztof Janowicz[a], Pascal Hitzler[b], Benjamin Adams[c], Dave Kolas[d], and Charles Vardeman II[e]

[a] *University of California, Santa Barbara, USA, e-mail: jano@geog.ucsb.edu*
[b] *Wright State University, Dayton, OH, USA, e-mail: pascal.hitzler@wright.edu*
[c] *The University of Auckland, New Zealand, e-mail: b.adams@auckland.ac.nz*
[d] *Raytheon BBN Technologies, MD, USA, e-mail: dkolas@bbn.com*
[e] *University of Notre Dame, Notre Dame, IN, USA, e-mail: cvardema@nd.edu*

Abstract. In 2010 Tim Berners-Lee introduced a 5 star rating to his Linked Data design issues page to encourage data publishers along the road to *good* Linked Data. What makes the star rating so effective is its simplicity, clarity, and a pinch of psychology – is **your** data 5 star? While there is an abundance of 5 star Linked Data available today, finding, querying, and integrating/interlinking these data is, to say the least, difficult. While the literature has largely focused on describing datasets, e.g., by adding provenance information, or interlinking them, e.g., by co-reference resolution tools, we would like to take Berners-Lee's original proposal to the next level by introducing a 5 star rating for Linked Data **vocabulary use**.

A Brief Motivation

Originally, the Linked Data design issues page introduced four rules on how to publish and interlink data in a human and machine readable way by using Semantic Web technologies [3]. While these rules are broad guiding principles, the provided explanations are fairly technical, e.g., discussing HTTP 303. Similarly, other authors have provided detailed best practice guidelines for publishing Linked Data [10,13].

Later, in 2010, Berners-Lee added a section about a 5 star rating for Linked (Open) Data [3]. This section has a different goal than the rest of the document and is largely non-technical. The goal is not to teach engineers how to create good Linked Data, instead it targets decision makers and stakeholders. It ties to *encourage*[1] (government) data owners to publish their data according to Linked Data principles by asking: *Is your Linked Open Data 5 Star?*. The first star is assigned for the big first step of making the data available on the Web (in whatever format). All following stars are intended to make the data easier to discover, use, and understand. The second star is assigned for making the data available in a machine readable, structured way (this can even be an Excel spreadsheet). The third star is for using non-proprietary formats (e.g., the Open Document Format; ODS instead of XLS). The fourth star is for using open W3C standards such as RDF to identify resources. Finally, the fifth star is for linking your own data to other datasets.

Again, the motivation here is to have a simple stack of consecutive steps that reward improved access to data by additional stars.

The Five Stars of Linked Data Vocabulary Use

So how *powerful*, *eas[y] for people to use* is 5 star Linked (Open) Data? Increasingly, we have come

[1] We will use *italics* type for direct quotes from [3].

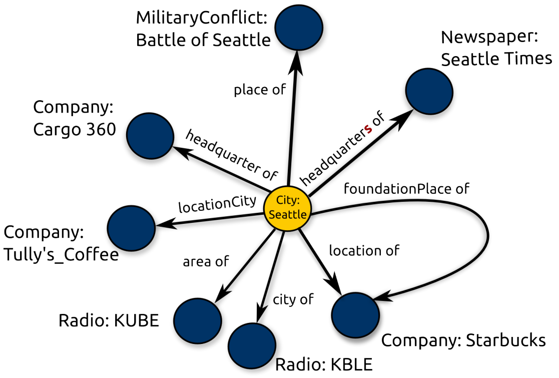

Fig. 1. Spatial DBpedia relations between the city of Seattle and other typed entities.

to believe that 5 star Linked Data is just the necessary **precondition** to what we really need. Just converting a CSV file to a set of RDF triples and linking them to another set of triples does not necessarily make the data more (re)usable to humans or machines. Various reasons/challenges and potential solutions have been discussed in the literature; e.g., see [2,8,12,14,17]. Recently, for instance, a major effort has been undertaken to standardize provenance information about Linked Data [15]. Here, however, we focus on a less active but not less important aspect: vocabularies. We use the term vocabulary here in a very broad and inclusive sense and do not further distinguish between schemata, ontologies, lightweight vocabularies, and so forth. Consequently, examples of vocabularies include FOAF,[2] DOLCE [7], schema.org, or the SSN-XG [6].

To give a concrete example, the statements

ex:HoratioNelson ex:diedIn ex:BattleOfTrafalgar.
ex:BattleOfTrafalgar rdf:type ex:NavalBattle.

are using a vocabulary if *ex* refers to a resource that states that

ex:NavalBattle rdfs:subClassOf ex:Battle.
ex:diedIn rdfs:subPropertyOf ex:participatedIn.

Similarly, stating that Horatio Nelson is of type foaf:Person is making use of the FOAF vocabulary.

Interestingly, the original 5 star rating does not make any assumptions about the use of vocabularies. In practice, however, querying Linked Data that do not refer to a vocabulary is difficult and understanding whether the results reflect the intended query is almost impossible. While there is a danger of over-engineering, a good vocabulary should restrict potential interpretations of the used classes and roles towards their intended meaning. Fig. 1, illustrates the fact that even in the presence of a lightweight ontology, querying even major Linked Data hubs, such as DBpedia [16], is challenging. How would you phrase a SPARQL query for companies located in Seattle?[3]

Instead of proposing new rules for engineering *good* vocabularies, we introduce 5 stars for Linked (Open) Data **vocabulary use** to *encourage* data owners, engineers, and practitioners to publish and use vocabularies on the Web. Similar to Berners-Lee's stars, which do not refer to the quality of the data as such (e.g., accuracy), our star rating is not concerned with the quality of the vocabularies (e.g., their consistency).

– ☆☆☆☆☆ **Linked Data without any vocabulary.** Zero stars are assigned to 5 star Linked Data that do not refer to any Web-accessible description of the used vocabulary.[4] For instance, 'LA temp 37.' may refer to a very cold night in the city of Los Angeles (in F) or a hot summer day (in C). LA is also the US postal abbreviation for Louisiana; so the statement could also refer to an average low temperature in the winter. Note that describing the dataset as being about 'temperatures of places' does not really help, i.e., using VoID [1] is an important first start but does not replace a good vocabulary.

– ★☆☆☆☆ **There is dereferencable human-readable information about the used vocabulary.** This information can be a detailed Web page documenting the vocabulary, a PDF file, a simple listing with selected examples, or even merely contact information. The intention here is similar to the second and third star in the original rating. We would also categorize controlled vocabularies here, as well as thesauri, as long as they are expressed in (controlled) natural language, graphics, tables, and so forth.

[2] http://www.foaf-project.org/

[3] Clearly, querying for all triples with roles that include Seattle and instances of Company is not a good solution.

[4] Though, visiting http://lov.okfn.org/dataset/lov/ may give you half a star ⯪ .

- ★★☆☆☆ **The information is available as machine-readable explicit axiomatization of the vocabulary.** Note that this is not restricted to a particular representation language. Ideally RDFS [5], OWL [11], RIF [4], or related W3C standards are used which are part of the Semantic Web stack. This is comparable to Berners-Lee's 4 star data rating. It is interesting to note that this is the level where SW **reasoning** comes into play, e.g., by exploiting the transitivity of certain roles, or inferring the type of an entity based on the defined domains and ranges.
- ★★★☆☆ **The vocabulary is linked to other vocabularies**. We believe that explicit alignments, e.g., via subClassOf or equivalentClass axioms, are often better than direct reuse of external vocabularies but both are acceptable. When working with data providers and software engineers, we often observe that they prefer to have control over their local vocabulary instead of importing a wide variety of (often under-specified, not regularly maintained) external vocabularies.[5] It is important to note that we refer to vocabulary-level links between classes and properties, not to links between individuals (e.g., via owl:sameAs).
- ★★★★☆ **Metadata about the vocabulary is available** (in a dereferencable and machine-readable form). This can be in form of the Ontology Metadata Vocabulary (OMV) [9], Vocabulary of a Friend (VOAF)[6], or other approaches. This can include information about the license model, contact person, last modification date, the used ontology language, the knowledge management methodology used to arrive at the vocabulary, and so forth.
- ★★★★★ **The vocabulary is linked to by other vocabularies.** Note that this is the reverse of the 3-star rating, i.e., the links point in the opposite direction. This is the only star on which the creator of the vocabulary has limited influence as it reflects the external usage and perceived usefulness. However, since 4 star Linked Data vocabularies are more powerful, easy to use than vocabularies with fewer stars, they will more likely earn the fifth star in the future. As explained before, these links have to be on the vocabulary-level.

Is your Linked Data vocabulary usage 5 star?

What's Next?

As a start, you could submit your next 5 star dataset and vocabulary to the Semantic Web journal's linked dataset descriptions or ontology description calls.[7]

References

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao, editors. *Describing Linked Datasets with the VoID Vocabulary.* W3C Interest Group Note 03 March 2011, 2011. Available from http://www.w3.org/TR/void/.

[2] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, 2013.

[3] T. Berners-Lee. Linked data: Design issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[4] H. Boley, G. Hallmark, M. Kifer, A. Paschke, A. Polleres, and D. Reynolds, editors. *RIF Core Dialect.* W3C Recommendation 22 June 2010, 2010. Available from http://www.w3.org/TR/rif-core/.

[5] D. Brickley and R. Guha, editors. *RDF Schema 1.1.* W3C Recommendation 25 February 2014, 2010. Available from http://www.w3.org/TR/rdf-schema.

[6] M. Compton, P. M. Barnaghi, L. Bermudez, R. Garcia-Castro, Ó. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. A. Henson, A. Herzog, V. A. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. R. Page, A. Passant, A. P. Sheth, and K. Taylor. The SSN ontology of the W3C semantic sensor network incubator group. *Journal of Web Semantics*, 17:25–32, 2012.

[7] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. In A. Gómez-Pérez and V. R. Benjamins, editors, *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Proceedings*, volume 2473 of *Lecture Notes in Computer Science*, pages 166–181. Springer, 2002.

[5] The same observation was recently highlighted by Oscar Corcho `bit.ly/1hKVj2F`.

[6] See `http://purl.org/vocommons/voaf`

[7] `http://semantic-web-journal.net/authors#types`

[8] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl:sameAs isn't the same: An analysis of identity in linked data. In P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web–ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer, Heidelberg, 2010.

[9] J. Hartmann, Y. Sure, P. Haase, R. Palma, and M. d. C. Suárez-Figueroa. OMV – Ontology Metadata Vocabulary. In *Proceedings of the International Workshop on Ontology Patterns for the Semantic Web, Galway, Ireland, November 2005*, 2005.

[10] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[11] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, editors. *OWL 2 Web Ontology Language: Primer*. W3C Recommendation 27 October 2009, 2009. Available at http://www.w3.org/TR/owl2-primer.

[12] P. Hitzler and F. van Harmelen. A reasonable semantic web. *Semantic Web*, 1(1):39–44, 2010.

[13] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[14] P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, and A. P. Sheth. Linked Data is Merely More Data. In D. Brickley, V. K. Chaudhri, H. Halpin, and D. McGuinness, editors, *Linked Data Meets Artificial Intelligence*, pages 82–86. AAAI Press, Menlo Park, CA, 2010.

[15] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, editors. *PROV-O: The PROV Ontology*. 2013. Available from http://www.w3.org/TR/prov-o/.

[16] J. Lehmann, R. ISele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 2014. to appear.

[17] A. Polleres, A. Hogan, A. Harth, and S. Decker. Can we ever catch up with the web? *Semantic Web*, 1(1):45–52, 2010.