Linked Data, Big Data, and the 4th Paradigm

Editorial

Pascal Hitzler, [a] Krzysztof Janowicz, [b]
[a] *Kno.e.sis Center, Wright State University, USA*
[b] *University of California, Santa Barbara, USA*

Around 2006, the inception of Linked Data [2] has led to a realignment of the Semantic Web vision and the realization that data is not merely a way to evaluate our theoretical considerations, but a key research enabler in its own right that inspires novel theoretical and foundational research questions. Since then, Linked Data is growing rapidly and is altering research, governments, and industry. Simply put, Linked Data takes the World Wide Web's ideas of global identifiers and links and applies them to (raw) data, not just documents. Moreover, and regularly highlighted by Tim Berners-Lee, ***A****nybody can say* ***A****nything about* ***A****ny topic* (AAA)[1] [1], which leads to a multi-thematic, multi-perspective, and multi-medial global data graph.

More recently, *Big Data* has made its appearance in the shared mindset of researchers, practitioners, and funding agencies, driven by the awareness that concerted efforts are needed to address 21st century data collection, analysis, management, ownership, and privacy issues. While there is no generally agreed understanding of what exactly is (or more importantly, what is *not*) Big Data, an increasing number of V's has been used to characterize different dimensions and challenges of Big Data: *volume*, *velocity*, *variety*, *value*, and *veracity*. Interestingly, different (scientific) disciplines highlight certain dimensions and neglect others. For instance, super computing seems to be mostly interested in the *volume* dimension while researchers working on sensor webs and the internet of things seem to push on the *velocity* front. The social sciences and humanities, in contrast, are more interested in *value* and *veracity*. As argued before [13,17], the *variety* dimensions seems to be the most intriguing one for the Semantic Web and the one where we can contribute most as a research community. Of course, these distinctions are not crisp and science is at its best when it aims at a holistic understanding. At the end, all V's have to be addressed in an interdisciplinary effort to substantially advance on the Big Data front [4].

The notion of variety as it applies to Big Data, of course, has a significant number of dimensions itself, including variety of data and representation formats as well as variety in terms of correctness, underlying conceptualizations or data models, temporal and spatial dependencies, etc. [13]. Untangling and understanding the important aspects of this variety notion is going to be part of the process of addressing them. Of course, variety also occurs in *Small Data*. However, in the case of small data volume or throughput (a.k.a., velocity), variety can usually be handled by established methods, e.g., by manual curation or explicit conversions. Thus, variety as a Big Data issue is distinct in that established small scale methods are insufficient. From this perspective, the Big Data notion of variety is a generalization of *semantic heterogeneity* as studied in the field of databases, artificial intelligence, Semantic Web, and cognitive science in general since many years.

The *4th Paradigm of Science* is yet another notion that has emerged within the last years and can be understood as the scientific view on how Big Data changes the very fabric of science [5]. With the omnipresence and availability of data from different times, locations, perspectives, topics, cultures, resolutions, qualities, and so forth, *exploration* becomes an additional (4th) paradigm of science. This raises *synthesis* to a new level. The study of relative (pixel) greenness in pictures extracted from public webcams to do research on phenology is just one example [10]. In other words, we can gain new insights by creatively combining what is already there – an idea that seems to align very well with *Linked* Data and Semantic Web technologies as drivers of integration. Interestingly, we

[1] Or AAAAA if you add space and time [12].

would argue that Big Data makes *small science* possible again as vasts amounts of data and processing power become available to individual scientists.

Summing up, it appears to be uncontroversial that Linked Data is part of the Big Data landscape. We would even go a bit further and claim that Linked Data is an ideal *testbed* for researching some key Big Data challenges and to experience the 4th paradigm in action.

Indeed, Linked Data reduces Big Data variability by some of the scientifically less interesting dimensions. For example, due to a general agreement on RDF [16] as basic data representation language for Linked Data, many syntactic issues vanish. Likewise, Linked Data relies on a relative small set of conventions, e.g., the use of vocabularies, and those vocabularies are created using a few formally well-defined languages (including, but not limited to OWL [7]). Even more, Linked Data can be accessed, stored, linked, queried, and so forth by a set of (largely) compatible free and open source tools and systems on regular hardware. Finally, most Linked Data are object centric (which reduces the mutli-mediality aspect). In this sense, Linked Data is a bit like Big Data *in a laboratory setting*, where certain variables are under control and thus can be ignored in the development of solutions or at least a deeper understanding of the issues. And once we have learned how to deal with the remaining variety dimensions in Linked Data, we are in a much better position to take further steps towards tackling Big Data at large.

In fact, it turns out that the variety challenges[2] which remain in Linked Data are still very substantial. As a particular example, just consider the case of integrated querying over multiple Linked Datasets. Of course, such integrated querying would be an obvious capability required for synthesis and exploration, and in particular such a capability would seem to be rather well aligned with the general promises of Semantic Web technologies. However, despite a good number of efforts and some advances (see, e.g., [14,15] and the references cited therein), the community still seems to be rather far away from a powerful, general, and practically feasible solution.

Such difficulties in making practical use of Linked Data are often attributed to *poor quality* of Linked Data [3,8,9,11,18]. In the context of the larger Big Data discussion, however, what may appear to be *quality* issues can be subsumed under the notion of *variety*: given the minimalistic agreed-upon requirements for Linked Data on the Web, and the fact that most datasets are community-created in a grassroots manner, it is only natural that there is not only a large variance in perspectives and underlying data models, but also a significant amount of genuine low-level quality issues such as erroneous and missing data, triplification errors, misleading *owl:sameAs* links, faulty syntax, and unavailable SPARQL endpoints.

Semantic Web and Linked Data researchers need to embrace these issues. They will not go away. Rather, we will need to find technical and methodological solutions which perform well even under such circumstances. And at the same time, we need to start creating a culture of best practices in data publishing which will alleviate some of the issues, and will thus make technical solutions easier to develop. Even more, following the previously outlined argumentation, it becomes clear that we need a way to communicate, illustrate, and document linked datasets in a way that is understandable to researchers and practitioners outside of the Semantic Web and computer/information science in general.

In this issue, we present the very first *Linked Dataset Description* papers which were accepted for publication in the Semantic Web journal. Initially, our call-for this new type of papers was intended as a one-time special issue. However the response was overwhelming and we received 27 submissions. We thus adopted dataset descriptions as a new standing paper type for the journal. Similarly to our argumentation for Tools & Systems papers [6], dataset descriptions are key *research enablers*. For instance, recent progress on foundational and theoretical aspects such as ontology alignment, semantic search, the combination of inductive and deductive approaches, and so forth would not have been possible without central Linked Data repositories such as DBpedia and Geonames, and their interlinkage.

By introducing a paper type for linked dataset descriptions we provide a publication outlet for these contributions, make them visible to a broader research community, archive them as notable contributions to development of the Semantic Web as research field, and finally allow for peer-review based quality control with a transparent set of review criteria. At the same time we give author's the chance to receive *academic credit* for their work. The peer-review aspect

[2] and, of course, also issues of volume, velocity, veracity, and value.

turned out to be an important fact as many of the initial submissions suffered from poor quality.[3] Over the last months we received feedback from editors, authors, and linked data enthusiasts that the existence of this paper type starts to affect the the quality of available linked datasets. As of May 2013 we received 59 linked dataset description papers,15 of these are presented in this issue.

Acknowledgements The first named author acknowledge support by the National Science Foundation under award 1017225 "III: Small: TROn – Tractable Reasoning with Ontologies." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] D. Allemang and J. Hendler. *Semantic web for the working ontologist: effective modeling in RDFS and OWL.* Morgan Kaufmann, 2008.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] 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 – 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2010.

[4] J. Hendler. Broad data: Exploring the emerging web of data. *Big Data*, 1(1):18–20, 2013.

[5] A. J. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA, 2009.

[6] P. Hitzler and K. Janowicz. Semantic web tools and systems. *Semantic Web*, 2(1):1–2, 2011.

[7] 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 from http://www.w3.org/TR/owl2-primer/.

[8] P. Hitzler and F. van Harmelen. A reasonable semantic web. *Semantic Web*, 1(1–2):39–44, 2010.

[9] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *3rd International Workshop on Linked Data on the Web (LDOW2010) at WWW2010, Raleigh, USA, April 2010*, 2010. Available from http://events.linkeddata.org/ldow2010/.

[10] N. Jacobs, W. Burgin, N. Fridrich, A. Abrams, K. Miskell, B. H. Braswell, A. D. Richardson, and R. Pless. The global network of outdoor webcams: properties and applications. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 111–120. ACM, 2009.

[11] 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.

[12] K. Janowicz. The role of space and time for knowledge organization on the semantic web. *Semantic Web*, 1(1-2):25–32, 2010.

[13] K. Janowicz and P. Hitzler. The digital earth as knowledge engine. *Semantic Web*, 3(3):213–221, 2012.

[14] A. Joshi, P. Jain, P. Hitzler, P. Yeh, K. Verma, A. Sheth, and M. Damova. Alignment-based querying of Linked Open Data. In R. Meersman, H. Panetto, T. Dillon, S. Rinderle-Ma, P. Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, and I. F. Cruz, editors, *On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012, Rome, Italy, September 10-14, 2012, Proceedings, Part II*, volume 7566 of *Lecture Notes in Computer Science*, pages 807–824. Springer, Heidelberg, 2012.

[15] V. Lopez, M. Fernández, E. Motta, and N. Stieler. PowerAqua: Supporting users in querying and exploring the Semantic Web. *Semantic Web*, 3(3):249–265, 2012.

[16] F. Manola and E. Miller, editors. *Resource Description Framework (RDF). Primer*. W3C Recommendation, 10 February 2004. Available from http://www.w3.org/TR/rdf-primer/.

[17] A. P. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *Int. J. Semantic Web Inf. Syst.*, 1(1):1–18, 2005.

[18] STKO. Place and location on the web of linked data. http://stko.geog.ucsb.edu/location_linked_data, 2013.

[3] http://blog.semantic-web.at/2012/08/09/whats-wrong-with-linked-data/