Key Ingredients for Your Next Semantics Elevator Talk

Krzysztof Janowicz[1] and Pascal Hitzler[2]

[1] Department of Geography, University of California, Santa Barbara, USA
jano@geog.ucsb.edu
[2] Kno.e.sis Center, Wright State University, Dayton, OH
pascal.hitzler@wright.edu

Abstract. 2012 brought a major change to the semantics research community. Discussions on the use and benefits of semantic technologies are shifting away from the *why* to the *how*. Surprisingly this more in stakeholder interest is not accompanied by a more detailed understanding of *what* semantics research is about. Instead of blaming others for their (wrong) expectations, we need to learn how to emphasize the paradigm shift proposed by semantics research while abstracting from technical details and advocate the added value in a way that relates to the immediate needs of individual stakeholders without overselling. This paper highlights some of the major ingredients to prepare your next *Semantics Elevator Talk*.

Keywords: Semantics, Ontology, Linked Data, Interoperability.

1 Introduction

Recently, we came across a Gartner Hype Cycle from 2006. It showed the term *Public Semantic Web* as currently entering the bottom of the *Trough of Disillusionment*, while *Corporate Semantic Web* was approaching the earlier *Peak of Inflated Expectations*. The Semantic Web community and related disciplines were questioning whether the field would recover or vanish. The Gartner picture made a dry statement: *5 to 10 years to mainstream adoption*. At hindsight, it seems amazing how profoundly accurate the forecast has turned out to be. Indeed, six years later, Steve Hamby announced 2012 as *The Year of the Semantic Web* in his Huffington Post article by listing a number of highly visible and prominent adoptions including Google's Knowledge Graph, Apple's Siri, Schema.org as cooperation between Microsoft, Google, and Yahoo!, Best Buy Linked Data, and so forth.[1] One could easily add more success stories for semantic technologies and ontologies such as the Facebook Open Graph protocol, The New York Times Web presence, or IBM's Watson system, and still just cover the tip of the iceberg.

[1] See http://www.huffingtonpost.com/steve-hamby/semantic-web-technology_b_1228883.html and www.huffingtonpost.com/steve-hamby/2012-the-year-of-the-sema_b_1559767.html

S. Castano et al.(Eds.): ER Workshops 2012, LNCS 7518, pp. 213–220, 2012.

While we see mainstream adoption in industry, academia, and governments, semantics research is far from over. Key research questions have yet to be solved and the wide adoption of more complex semantic technologies and of knowledge engineering is a distant goal on the horizon. Often, past research has provided conceptual insights and purely theoretical approaches to pressing topics such as how to address semantic interoperability, but failed to deliver ready-made solutions. As a research community, we are suddenly faced with discussions shifting away from the *why* to the *how*. Our technical language, loaded with the infamous three-letter acronyms, is not suitable to explain the immediate added value of adopting semantic technologies to stakeholders. With the dawning data revolution, the Semantic Web community is confronted with the need to provide working solutions for data publishing, retrieval, reuse, and integration in highly heterogeneous environments. Interdisciplinary science and knowledge infrastructures such as NSF's Earthcube[2] are among the most promising areas to put semantics to work and to show the immediate added value of our research [1].

Targeting the semantics research community, this paper highlights some of the ingredients required to prepare a semantics elevator talk that explains the value proposition of the Semantic Web to interdisciplinary scientists and at the same time circumnavigates common misunderstandings about the adoption of semantic technologies.

2 The Value Proposition of the Semantic Web

What can be achieved by using the Semantic Web that was not possible before is among the most frequent questions raised when introducing the Semantic Web to stakeholders, and *nothing* is probably the most honest answer. Instead, and more appropriately, one should ask whether a certain project would be realized *at all* without the aid of semantic technologies – in other words, the question is not what is doable, but what is *feasible*. In the following, we list three examples that demonstrate the added value of semantics in different stages of scientific workflows, and which are driven by the immediate needs of scientists instead of abstract assertions.

2.1 Publishing and Retrieving

Participating in the Semantic Web is a staged process and the entry level has been constantly lowered over the past few years, thereby contributing to the success of Linked Data [2] in science, governments, and industry. For the individual scientist, the added value of semantic technologies and ontologies starts with publishing own data. By creating more intelligent metadata, researchers can support the discovery and reuse of their data as well as improve the reproducibility of scientific results. This aspect is increasingly important as journals and conferences ask authors to submit their data along with the manuscripts.

[2] `http://earthcube.ning.com/`

Semantically annotated data also enables search beyond simple keyword matching. Google's *things not strings* slogan implemented in their new Knowledge Graph shows semantic search in action and highlights how single pieces of data are combined and interlinked flexibly.[3] In a scientific context and combined with Big Data, semantic search and querying will go further and allow to answer complex scientific questions that span over scientific disciplines [1]. With EarthCube, NSF is currently establishing such an integrated data and service infrastructure across the geosciences. New semantics-enabled geographic information retrieval paradigms employ ontologies to assist users in browsing and discovering data based on analogies and similarity reasoning [3,4,5]. To give concrete examples, the paradigm shift from data silos to interlinked and open data will support scientists in searching for appropriate study areas, in finding data sources which offer a different perspective on the same studied phenomena to gain a more holistic view, and in interlinking their own data with external datasets instead of maintaining local and aging copies.

2.2 Interacting and Accessing

One of the key paradigm shifts proposed by the Semantic Web is to enable the creation of smart data in contrast to smart applications. Instead of developing increasingly complex software, the so-called business logic should be moved to the (meta)data. The rationale is that smart data will make all future applications more usable, flexible, and robust, while smarter applications fail to improve data along the same dimensions. To give a concrete example, faceted search interfaces and semantics-enabled Web portals can be created with a minimum of human interaction by generating the facets via the roles and their fillers from the ontologies used to semantically annotate the data at hand. Changes in the underlying ontologies and the used data are automatically reflected in the user interface. In fact, users can even select their preferred Linked Data browser as along as the data is available via a SPARQL endpoint. One example for such a semantics-enabled portal that is semi-automatically generated out of ontologies and data is the Spatial Decision Support portal [6]. In terms of added value, semantic technologies and ontologies reduce implementation and maintenance costs and enable users to access external datasets via their preferred interface, thus benefiting data publishers and consumers. Due to the high degree of standardization and reasoning capabilities enabled by the formal semantics of knowledge representation languages, most available Semantic Web software is compatible. For instance, data can be easily moved between triple stores.

2.3 Reusing and Integrating

Semantic technologies and ontologies support horizontal and vertical workflows, i.e., they offer approaches for all phases starting from data publishing, sharing,

[3] `http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html`

discovery, and reuse, to the integration of data, models, and services in heterogeneous environments. For many scientists and engineers, the reuse and integration aspects may be those with the clearest added value, as 60% of their time is spent on making data and models compatible [7]. By restricting the interpretation of domain vocabularies towards their intended meaning, ontologies reduce the risk of combining unsuitable data and models. A purely syntactic approach or natural language descriptions often fail to uncover hidden incompatibilities and may result in misleading or even wrong results [8].

However, improving semantic interoperability is not the only added value with respect to data reuse and integration. Semantic technologies also support the creation of rules for integrity constraint checking. To give a concrete example, a scientist may import vector data on afforested areas into a semantics-enabled Geographic Information System that checks the data against a selected ontology to display those areas that correspond to a specific *Forest* definition [9]. Finally, semantic technologies and ontologies can also assist scientists in selecting appropriate analysis methods, e.g., by verifying that a particular statistics returns meaningful results when applied to the dataset at hand.

3 Adoption Steps

For potential adopters of semantic technologies, it is often important that rapid progress is made which quickly leads to visible and testable added value. This aspect should not be underestimated. Adopters need to justify their investments, and it could be perceived as a high risk approach if benefits were a long time coming. At the same time, the powerful added value of adopting semantic technologies only unfolds in full in later stages of adoption. The challenge is, thus, to keep the ball rolling through the early adoption stages, such that the greater benefits can be reaped in the medium and long term. The need for rapid adoption can be met with semantic technologies, however a certain minimum of care needs to be taken to make sure that adoption reaches the later and even more beneficial stages. In this section, we point out some key issues related to this staged adoption.

At first, however, it is important for adopters to realize that some semantic technologies have a steep learning curve, and, similarly to engineering disciplines, require a certain routine. Adopters will need an infusion of expert knowledge, either by hiring semantic technology experts or by closely cooperating with them. These experts should be honest about the limits of certain technologies and willing to listen to domain and application problems instead of approaching them with domain-independent blueprints. The Semantic Web is extremely rich, there is always more than one way to go. However, this also requires that potential adopters communicate their needs and ask about the pros and cons of available options. All these problems are well known from working in interdisciplinary teams and, at its core, semantics is all about heterogeneity.

3.1 Rapid Initial Adoption

Rapid adoption starts with publishing data following the Linked Data paradigm. In essence, this means making the data available in a standardized and simple syntactic format, namely in RDF [10]. It is important to understand that this first step does not necessarily add any relevant semantics to the data.

Immediate benefits for the adopter include the following.

- Stakeholders can find the data and access it with common tools which can handle RDF and the RDF semantics. Hence, the barrier to find and reuse data is lowered considerably.
- The adopter's data will become part of the active research community which is concerned with analyzing, understanding, improving, interlinking, and using Linked Data for various purposes.
- Data can be combined with external data via links without the need to keep local copies of such external datasets.
- The adopter gains visibility and reputation by contributing to an open culture of data and as part of the state-of-the-art Linked Data effort.

With those benefits in mind, it is also important to point out what Linked Data does *not* deliver [11,12,13].

- A common syntax helps to lower the barrier for reuse, but does not address semantic interoperability nor does it enable complex queries across datasets, which means that data curation is still a major and non-trivial effort. Essentially, data that is published using informal or semi-formal vocabularies is still wide open to ambiguities and misinterpretations. While this may be less problematic for interaction with human users, it sets clear limits for software agents.
- The *links* in Linked Data are often created ad-hoc with a more-is-better mentality instead of strategies to assess quality, or to maintain and curate already established links. Indeed, many of those links are `owl:sameAs` links which, however, are usually not meant to carry the formal semantics they would inherit from the Web Ontology Language OWL [11,14].
- The paradigm shift to triples as units of meaning and URIs as global identifiers alone is not sufficient to contribute to the Linked Data cloud. A set of methods and tools is required [15]. As research community we have to provide best practice and strategies for different types of stakeholders and projects.

Summing up, publishing Linked Data is a major first step and offers immediate added value at low cost (in terms of time and infrastructure). This step alone, however, does not automatically enable many of the promises of the Semantic Web. In fact, many of the early Linked Data projects merely ended up as more data [13].

3.2 Medium- and Long-Term Bootstrapping

In order to understand how to initiate a medium- and long-term process in adopting *deep* semantic technologies, let us first dwell on one of the key fallacies

to adopting semantics in a *rapid* fashion. As pointed out above, such a rapid adoption essentially establishes a common syntax and otherwise relies on the use of vocabularies whose meaning is usually not formally defined and requires substantial human interaction and interpretation.

To make a very simple example for potential difficulties, consider the ad-hoc vocabulary term `ex:hasEmail`, informally described as an RDF propoerty having as values strings which are email addresses of contact persons of a particular nature preserve. Now assume that some of these contact persons use a common email account, e.g., to share responsibilities. Usually, this does not cause any difficulties and is, in fact, common practice. However, a knowledge engineer may, at some later stage, be in need of having more powerful semantics at hand, e.g., because on the Web email addresses are often used as identifiers for account holders, and thus it seems reasonable to assume that `ex:hasEmail` is an inverse functional property in the exact sense in which OWL specifies it.[4] Regretfully, it turns out that this apparently harmless strengthening of the semantics of the vocabulary term `ex:hasEmail` now yields undesired consequences. According to the OWL semantics, we can now conclude that all contact persons having the same email address are, in fact, identical (in the sense of `owl:sameAs`). This introduces many undesirable logical consequences and may contradict with existing schema knowledge. Such problems are even more likely when reusing existing ontologies that do not provide a clear maintenance and evolution strategy as well as by being too careless with the use of `owl:sameAs` links to external (and fluid) datasets.

The problem lies in the attempt to strengthen the semantics of previously under- or informally specified vocabulary terms used to semantically enable data. This is especially problematic for large datasets from different sources that were created and maintained by different parties. In many cases a retroactive "deep semantification" will be difficult or even impossible if it has not been introduced up front.

There is no simple solution for this issue, and a *rapid adoption* approach will sooner or later always lead to such difficulties, semantic aging being another example [17]. At this stage, i.e., to strengthen the semantics of vocabularies, considerable effort will have to be invested in curating the data by mapping it to more expressive ontologies. Regretfully, provenance information for data may already be missing, so that a curation of the data will not always be feasible. In the end there is a trade-off between rapid adoption and ease of establishing deep semantics capabilities, which has to be considered for each use case and application area.

However, some of the overhead work can be avoided by treading carefully from the start. It helps to reuse existing high-quality ontologies and ontology design paterns, and it is important to have a clear understanding of the formal semantics of the adopted ontology language (e.g., OWL), and its implications, even if the initial plan is to only use simple language constructs. To give another elementary example, novices in conceptual modeling often confuse class

[4] FOAF [16] treats email addresses this way, for example.

hierarchies with partonomies, and may be tempted to use `rdfs:subClassOf` as a part-of relationship. The same is true for the more informally used is-a and instance-of relations. By having a clear grasp of the formal semantics of OWL (and RDFS) vocabulary, such mistakes can be avoided.

Summary

We have presented some key aspects concerning the elevation of semantic technologies for adoption in the sciences. In particular, we discussed central value propositions of semantic technologies and ontologies as well as potential roadblocks related to their adoption. While we are aware that the presented list of topics is incomplete and only outlined here, we hope that it will help to start a discussion on how to clarify the value proposition of the Semantic Web within the sciences, communicate paradigm changes and not technologies, lay out roadmaps for knowledge infrastructures such as NSF's EarthCube, and foster our shared visions without overselling them.

Acknowledgement. The second author acknowledges 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. Janowicz, K., Hitzler, P.: The Digital Earth as a knowledge engine. Semantic Web Journal (to appear, 2012), `http://www.semantic-web-journal.net/`
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Jones, C.B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 322–335. Springer, Heidelberg (2001)
4. Nedas, K., Egenhofer, M.: Spatial-scene similarity queries. Transactions in GIS 12(6), 661–681 (2008)
5. Janowicz, K., Raubal, M., Kuhn, W.: The semantics of similarity in geographic information retrieval. Journal of Spatial Information Science (2), 29–57 (2011)
6. Li, N., Raskin, R., Goodchild, M., Janowicz, K.: An ontology-driven framework and web portal for spatial decision support. Transactions in GIS 16(3), 313–329 (2012)
7. NASA: A.40 computational modeling algorithms and cyberinfrastructure (December 19, 2011). Technical report, National Aeronautics and Space Administration (NASA) (2012)
8. Kuhn, W.: Geospatial Semantics: Why, of What, and How? In: Spaccapietra, S., Zimányi, E. (eds.) Journal on Data Semantics III. LNCS, vol. 3534, pp. 1–24. Springer, Heidelberg (2005)

9. Lund, G.: Definitions of forest, deforestation, afforestation, and reforestation. [online] gainesville, va: Forest information services. Technical report (2012), available from the world wide web: http://home.comcast.net/~gyde/DEFpaper.htm
10. Manola, F., Miller, E.: RDF primer, W3C Recommendation. Technical report, W3C, February 10 (2004)
11. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
12. Hitzler, P., van Harmelen, F.: A reasonable Semantic Web. Semantic Web 1(1-2), 39–44 (2010)
13. Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P.: Linked Data is Merely More Data. In: AAAI Spring Symposium Linked Data Meets Artificial Intelligence, pp. 82–86. AAAI Press (2010)
14. Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. (eds.): OWL 2 Web Ontology Language: Primer. W3C Recommendation, October 27 (2009), http://www.w3.org/TR/owl2-primer/
15. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)
16. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.98. Namespace Document, August 9 (2010), http://xmlns.com/foaf/spec/
17. Schlieder, C.: Digital heritage: Semantic challenges of long-term preservation. Semantic Web 1(1-2), 143–147 (2010)