

Semantic Web Research anno 2006: main streams, popular fallacies, current status and future challenges

Frank van Harmelen

Vrije Universiteit Amsterdam
Frank.van.Harmelen@cs.vu.nl

Abstract. In this topical¹ paper we try to give an analysis and overview of the current state of Semantic Web research. We point to different interpretations of the Semantic Web as the reason underlying many controversies, we list (and debunk) four false objections which are often raised against the Semantic Web effort. We discuss the current status of the Semantic Web work by reviewing the current answers to four central research questions that need to be answered, and by surveying the uptake of Semantic Web technology in different application areas. Finally, we try to identify the main challenges facing the Semantic Web community.

1 Which Semantic Web?

It has already been pointed out by Marshall and Shipman in [1] that the term “Semantic Web” is used to describe a variety of different goals and methods. They distinguish (1) the Semantic Web as a universal library for human access; (2) as the habitat for automated agents and web-services²; and (3) as a method for federating a variety of databases and knowledge bases. And although we in no way share their rather pessimistic analysis of the possibilities for each of these three scenario’s (founded as they are on rather strawman versions of each of them), we do agree that it is important to unravel the different ambitions that underly the “Semantic Web” term.

In the current Semantic Web work, we distinguish two main goals. These goals are often unspoken, but the differences between them often account for many debates on design choices, on the applicability of various techniques, and on the feasibility of applications.

¹ in the sense of: “of current interest”, “concerning contemporary topics of limited validity”

² although Marshall and Shipman do not actually use the term web-services

Interpretation 1: The Semantic Web as the Web of Data

In the first interpretation (close to Marshall and Shipman's third option), the main aim of the Semantic Web is to enable the integration of structured and semi-structured data-sources over the Web. The main recipe is to expose datasets on the web in RDF format, to use RDF Schema to express the intended semantics of these data-sets, in order to enable the integration and unexpected re-use of these data-sets.

A typical use-cases for this version of the Semantic Web is the combination of geo-data with a set of consumer ratings for restaurants in order to provide an enriched information source.

Interpretation 2: The Semantic Web as an enrichment of the current Web

In the second interpretation, the aim of the Semantic Web is to improve the current World Wide Web. Typical use-cases here are improved search engines, dynamic personalisation of web-sites, and semantic enrichment of existing web-pages.

The source of the required semantic meta-data in this version of the Semantic Web is mostly claimed to come from automatic sources: concept-extraction, named-entity recognition, automatic classification, etc. More recently, the insight is gaining ground that the required semantic markup can also be produced by social mechanisms of in communities that provide large-scale human-produced markup.

Of course there are overlaps between these two versions of the Semantic Web: they both rely on the use of semantic markup, typically in the form of meta-data described by ontology-like schemata. But perhaps more noticeable are the significant differences: different goals, different sources of semantics, different use-cases, different technologies.

2 Four popular fallacies

The Semantic Web is subject to a stream of strongly and often polemically voiced criticisms³. Unfortunately, not all of these are equally well informed. A closer analysis reveals that many of these polemics attribute a number of false assumptions or claims to the Semantic Web programme. In this section we aim to identify and debunk these fallacies.

Fallacy 1: The Semantic Web tries to enforce meaning from the top

This fallacy claims that the Semantic Web, enforces meaning on users through its standards OWL and RDF(S). The repost to this fallacy is easy. The only

³ e.g. <http://www.shirky.com/writings/semantic-syllogism.html> and <http://www.csd1.tamu.edu/~marshall/mc-semantic-web.html>

meaning that OWL and RDF(S) enforce is the meaning of the connectives in a language that users can use to express their own meaning. The users are free to choose their own vocabulary, and to assign their own meaning to terms in this vocabulary, to describe whatever domain of their choice. OWL and RDF(S) are entirely neutral in this.

The situation is comparable to HTML: HTML does not enforce the lay-out of web-pages “from the top”. All HTML enforces is the language that people can use to describe their own lay-out. And HTML has shown that such an agreement on the use of a standardised language (be it HTML for the lay-out of web-pages, or RDF(S) and OWL for their meaning) is a necessary ingredient for world-wide interoperability.

Fallacy 2: The Semantic Web requires everybody to subscribe to a single predefined meaning for the terms they use.

Of course, the meaning of terms cannot be predefined for global use. Of course, meaning is fluid and contextual. The motto of the Semantic Web is not the enforcement of a single ontology. It’s motto is rather “let a thousand ontologies blossom”. That is exactly the reason why the construction of mappings between ontologies is such a core topic in the Semantic Web community (see [2,3,4] for some surveys). And such mappings are expected to be partial, imperfect and context-dependent.

Fallacy 3: The Semantic Web will require users to understand the complicated details of formalised knowledge representation.

Indeed some of the core technology of the Semantic Web relies on intricate details of formalised knowledge representation. The semantics of RDF Schema and OWL, and the layering of the subspecies of OWL are difficult formal matters. The design of good ontologies is a specialised area of Knowledge Engineering. But for most of the users of (current and future) Semantic Web applications, such details will be entirely “under the hood”, just as the intricacies of CSS and (X)HTML are under the hood of the current Web. Navigation or personalisation engines can be powered by underlying ontologies, expressed in RDF Schema or OWL, without the user ever being confronted with the ontology, let alone its representation language.

Fallacy 4: The Semantic Web people will require the manual markup of all existing web-pages

It’s hard enough for most web-site owners to maintain the human-readable content of their site. They will certainly not maintain a second parallel version in which they will have to write a machine-accessible version of the same information in RDF or OWL. If this were the case, that would indeed spell bad news for the Semantic Web. Instead, Semantic Web applications rely on large-scale

automation for the extraction of such semantic markup from the sources themselves. This will often be very lightweight semantics, but for many applications, that has shown to be enough.

Notice that this fallacy mostly affects interpretation 2 of the Semantic Web (previous section), since massive markup in the “Web of data” is much easier: the data is already available in (semi-)structured formats, and is often already organised by database schema’s that can provide the required semantic interpretation.

3 Current status

In this section, we will briefly survey the current state of work on the Semantic Web in two ways. First we will try to assess the progress that has been made in answering four key questions on which the success of the Semantic Web relies. Secondly, we will give a quick overview of the main areas in which Semantic Web technology is currently being adopted.

3.1 The Four Main Questions

Question 1: where does the meta-data come from?

As pointed out in our Fallacy No. 4, much of the semantic meta-data will have to come from Natural Language Processing and Machine Learning technology. And indeed, these technologies are delivering this promise. It is now possible with off-the-shelf technology to produce semantic markup for very large corpora of web-pages (millions of pages) by annotating them with terms from very large ontologies (hundreds of thousands of terms), at a sufficiently quality precision and recall to drive semantic navigation interfaces. Our own work on the DOPE prototype is only one of many examples that can be given: 5 million web-pages indexed with an ontology of 235.000 concepts, used for query disambiguation, narrowing, widening and semantic clustering of query results [5].

More recently (and for many in the Semantic Web community somewhat unexpected) is the capability of social communities to do exactly what Fallacy 4 claims is impossible: providing large amounts of human-generated markup. Millions of images, hundreds of millions of manually provided with meta-data tags on some of the most popular “Web 2.0” sites.

Question 2: where do the ontologies come from?

As pointed out by [6], the term *ontology* as used by the Semantic Web community now covers a wide array of semantic structures, from lightweight hierarchies such as MeSH⁴ to heavily axiomatised ontologies such as GALEN⁵.

⁴ <http://www.nlm.nih.gov/mesh/>

⁵ <http://www.opengalen.org/>

The lesson of a decade worth of Knowledge Engineering and half a decade of Semantic Web research is that indeed the world is full of such “ontologies”: companies have product catalogues, organisations have internal glossaries, scientific communities have their public meta-data schemata. These have typically been constructed for other purposes, most often pre-dating the Semantic Web, but very useable as material for Semantic Web applications.

There are also significant advances in the area of ontology-learning, although results there remain mixed: obtaining the concepts of an ontology is feasible given the appropriate circumstances, but placing them in the appropriate hierarchy with the right mutual relationships remains a topic of active research.

Question 3: what to do with many ontologies?

As stated in our rebuttal to fallacy No. 2, the Semantic Web crucially relies on the possibility to integrate multiple ontologies. This is known as the problem of ontology alignment, ontology mapping or ontology integration, and is indeed one of the most active areas of research in the Semantic Web community. Excellent surveys of the current state of the art are provided by [2,3,4].

A wide array of techniques is deployed for solving this problem, with ontology mapping techniques based on natural language technology, based on machine-learning, on theorem-proving, on graph-theory, on statistics, etc.

Although encouraging results are obtained, this problem is by no means solved, and automatically obtained results are not yet good enough in terms of recall and precision to drive many of the intended Semantic Web use-cases. Consequently, ontology-mapping is seen by many as the Achilles Heel of the Semantic Web.

Question 4: wheres the “Web” in the Semantic Web?

The Semantic Web has sometimes been criticised as being too much about “semantic” (i.e. large-scale distributed knowledge-bases), and not enough about “web”. This was perhaps true in the early days of Semantic Web developments, where there was a focus on applications in rather circumscribed domains like intranets. This initial emphasis is still visible to a large extent: many of the most successful applications of Semantic Web technology are indeed on company intranets. Of course the main advantage of such intranet-applications are that the ontology-mapping problem can to a large extent be avoided.

Recent years have seen a resurgence in the Web-aspects of Semantic Web applications. A prime example of this is the deployment of FOAF technology⁶, and of semantically organised P2P systems (see e.g. the collection of work in [7]).

Of course the Web is more than just textual documents: non-textual media such as images and videos are an integral part of the Web. For the application of Semantic Web technology to such non-textual media we must for the foreseeable

⁶ <http://www.foaf-project.org/>

future rely on human-generated semantic markup (as discussed above), given the difficulty of automatically generating meaningful markup for such media.

Main application areas

It is beyond the scope of this brief paper to give an in-depth and comprehensive overview of all Semantic Web applications. We will limit ourselves to a bird's eye survey.

Looking at industrial events either dedicated events⁷ or co-organised with the major international scientific Semantic Web conferences, we observe the following.

A healthy uptake of Semantic Web technologies is beginning to take shape in the following areas:

- knowledge management, mostly in intranets of large corporations
- data-integration (Boeing, Verison and others)
- e-Science, in particular the life-sciences⁸
- convergence with Semantic Grid

If we look at the profiles of companies active in this area, we see a distinct transition from small start-up companies such as Aduna, Ontoprise, Network Inference, Top Quadrant (to name but a few) to large vendors such as IBM (their Snobase ontology Management System⁹, HP (with their popular Jena RDF platform¹⁰, Adobe (with their RDF-based based XMP meta-data framework) and Oracle (now lending support for RDF storage and querying in their prime data-base product).

However, besides the application areas listed above, there is also a noticeable lack of uptake in some other areas. In particular, promises in the areas of

- personalisation,
- large-scale semantic search (i.e. on the scale of the World Wide Web, not limited to intranets),
- mobility and context-awareness

are largely unfulfilled.

A pattern that seems to emerge between the succes unsuccessfull application areas is that the succesfull areas are all aimed at closed communities (employees of large corporations, scientists in a particular area), while the applications aimed at the general public are still in the laboratory phase at best. The underlying reason for this could well be as discussed above, namely the difficulty of the ontology mapping.

⁷ e.g. <http://www.semantic-conference.com/>

⁸ see e.g. <http://www2006.org/speakers/stephens/stephens.ppt> for some state-of-the-art industrial work

⁹ <http://www.alphaworks.ibm.com/tech/snobase>

¹⁰ <http://jena.sourceforge.net/>

4 Challenges

Many of the challenges that we outlined in an earlier paper [8] are in the mean-time active areas of research:

- scale (with inference and storage technology are now scaling to the order of billions of RDF triples,
- ontology evolution and change
- ontology mapping, as outlined above.

However, a number of items on the research agenda are hardly tackled, but do have a crucial impact on the feasibility of the Semantic Web vision. In particular:

- the mutual interaction between machine-processable representations and the dynamics of social networks of human users
- mechanisms to deal with trust, reputation, integrity and provenance in a semi-automated way
- inference and query facilities that are sufficiently robust to work in the face of limited resources (be it either computation time, network latency, memory or storage-space), and that can make intelligent trade-off decisions between resource use and output-quality

References

1. Marshall, C.C., Shipman, F.M.: Which semantic web? In: *HYPERTEXT '03: Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, ACM Press (2003) 57–66
2. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review Journal (KER)* (18(1)) (2003) 1–31
3. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. *The VLDB Journal* (10(4)) (2001) 334–350
4. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. *Journal on Data Semantics (IV)* (2005) 146–171
5. Stuckenschmidt, H., van Harmelen, F., de Waard, A., Scerri, T., Bhogal, R., van Buel, J., Crowlesmith, I., Fluit, C., Kampman, A., Broekstra, J., van Mulligen, E.: Exploring large document repositories with rdf technology: The dope project. *IEEE Intelligent Systems* **19**(3) (2004) 34–40
6. Jasper, R., Uschold, M.: A framework for understanding and classifying ontology applications. In: *Proceedings 12th Int. Workshop on Knowledge Acquisition, Modelling, and Management KAW*. (1999) 4–9
7. Staab, S., Stuckenschmidt, H.: *Semantic Web and Peer-to-peer: Decentralized Management and Exchange of Knowledge and Information*. Springer (2005)
8. van Harmelen, F.: How the semantic web will change kr: challenges and opportunities for a new research agenda. *The Knowledge Engineering Review* **17**(1) (2002) 93–96