

The Semantic Web 10th Year Update

James Hendler
Tetherless World Chair
Dept of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
hendler@cs.rpi.edu

ABSTRACT

This talk is based on the fact that the Scientific American article by Tim Berners-Lee, Ora Lassila and James Hendler was published on May 21, 2001. The WIMS'11 conference is an appropriate event to reflect on the 10 years - successes and misses.

Keywords

Web, Linked Data, Semantic Web, Open Data

I spend a lot of my time fielding questions from various people on what the Semantic Web is (and isn't) and about its status – so I am writing this paper in the form of an FAQ. Below are some of the questions I get asked the most, and my answers.

Q. When will we see the Semantic Web emerge? (this question is also asked as "with all the hype, why haven't we seen anything)

A. My answer to this one is that it shows a certain ignorance of what the Semantic Web is all about. In particular, the Semantic Web is primarily an infrastructure technology that will bring new information to the Web, especially structured and semi-structured materials), that will help link sites to each other (via links through the "semantic space" - see below), and that will help create more information that can be, in some simple sense of the term, "understood" by the computer. Note that these things must be exposed through integration with the current Web -- so that the revolutionary capabilities offered by the Semantic Web (and I still do believe it will provide a revolution in capabilities) will primarily be seen through improvements in functionality to existing web sites and to new Web applications - but they'll still be deployed through the browser and look like the other stuff on the Web.

So properly asked, the question is when will Semantic Web technologies be deployed widely on the Web. I think that is a

more interesting question. We know that a lot of stuff is out there, and many sites use the maturing RDF already -- examples include the new Yahoo! Web home site and a number of their other pages, Google's Rich Snippets and Yahoo's search monkey, which use RDFa, freebase.com (acquired by Google), and others that either use of export information in SW formats, etc. -- there's a good presentation on this that Frank van Harmelen prepared¹). In the past year we have also seen Facebook adopt an RDFa framework that can be used for "like buttons," and which they report are used by 10-15% of the like buttons you see on the Web. That's a lot of semantics out there!

So basically, I think Web 3.0 is here, but the Web is so big that you don't always see it. More importantly, this is really more of an infrastructure technology, so without a "Web 3.0 Inside" sticker of some kind, you don't know some of the sites you already use are using it. When John Markoff wrote a NY Times article more or less coining the term Web 3.0² in November 2006, he was responding to a panel that had three new companies on it – MetaWeb (now part of Google), RadarNetworks (acquired by ESRI) and Powerset (brought out by Microsoft). Powerset was bought by Microsoft, and their technology is now reputed to be a significant contributor to bing.com, the new "decision engine" being promoted by Microsoft. Google supports something called "rich snippets" and Yahoo! something called "Search monkey" both of which expose semantic technologies in interesting ways. Several of the larger social networking sites are reputed to be working with the search engine companies to exploit these technologies, which has aroused a lot of interest in this community. So much is happening right now, but a lot of it is under the hood - not hard to find, but you have to know how to look.

The key thing about the above is that on the Web, once something starts to prove successful, it tends to grow. Every year since 2001, we've seen more and more of this below the hood stuff happening, but now with these sorts of breakthroughs, we're really seeing things starting to heat up. I could claim it is already used daily by those on the Web, but I'd be exaggerating as probably not more than a few hundred thousand of the Web's billion users hit these sites on a regular basis. But that will grow rapidly in the next few years and I feel very comfortable saying we're going to see a staggering amount of this stuff coming along. So I'd be surprised

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WIMS'11, May 25-27, 2011 Sogndal, Norway

Copyright © 2011 ACM 978-1-4503-0148-0/11/05... \$10.00.

¹ <http://www.slideshare.net/Frank.van.Harmelen/semantic-web-good-news>

² http://www.nytimes.com/2006/11/12/business/12web.html?_r=1

if it's not the case in the next five years that more than 10% of Web users will regularly be hitting an application with some sort of semantic web technologies involved (and some large percentage of that will be using Semantic Web apps on a regular basis).

So my real answer to this question is that it is here, and it is growing, but that we still have plenty of space to grow. Exciting things are coming, and will continue to do so.

(However, I caution to add that the technologies seeing the biggest use now are those that were the first ones out of the research laboratories. There's a lot more interesting things coming, but the day of "AI" and "intelligent agents" is still not going to be here in five or ten years, but more and more of it will be coming each year, so it's an exciting time to be in this area).

Q. Is Semantic Search part of the Semantic Web (same question for Semantic social networks, semantic match, semantic ?x for most ?x)?

A. This one is harder than it might seem. The question is whether "AI on the Web" is inherently Semantic Web, and I think the answer to that is a clear no. There's many things that use various models of learning across large statistical datasets, evolutionary or other techniques are used on "human computation" Web sites, and so too are other AI techniques deployed on the Web. In most of these cases, however, the missing part is linking - the Web stuff - lots of techniques will make the Web better, but the real key is how the links work.

But a lot of these systems are starting to use Semantic Web technologies and URIs within the application. Those are harder to answer. For example, Powerset (now part of Microsoft) was reputed to use Wordnet and various ontologies - whether those were explicitly in Semantic Web formats or not, is unclear, but it was clear that the creation of these things, the import and export of information was related to Semantic Web formats etc.? So is this Semantic Web? I think so, because we are using the ontological stuff in a way that, eventually, we'll be able to use and link to other people's stuff. On the other hand, it's currently in a single app - so I'd say we're talking "Semantic IntraWeb" - an idea that is not well defined, but seems to be where a lot of the "web 3.0" players are right now.

Q. So what is it that your research group at Rensselaer is doing?

The field of AI has generally stressed expressive knowledge representation (being able to say something like "a hand that belongs to a human has five fingers, one of which is the thumb") or on having lots of data and no knowledge at all (as the many machine learning projects currently being deployed). However, a small amount of knowledge applied to a large dataset seems like an extremely important, and largely ignored, area of Web development. My research is now looking at things like "very scalable" reasoning and also on "data on demand" systems - that is, in many applications there is so much data that it cannot be easily stored in a local machine (for example in science applications where we now see petabytes of data). We are looking at technologies that could, on the fly, find and merge appropriate pieces of very large datasets into custom "data caches" and make those available in Web applications. The key to a lot of this is that

being able to scale these things requires some semantics, but not the traditional KR that AI people have explored nor can you use only the relational model that has been the hallmark of database research.

Q. How about describing one or two cool projects that you are focused on now and that you believe will lead to promising developments in this area in the near term?

Here's two I'm particularly excited about right now:

In <http://logd.tw.rpi.edu> we've been taking the data that the US government has been releasing in the Data.gov project and making it available in Semantic Web formats. This allows us to rapidly create visualizations, link it to other datasets (either from there or other govt sources), and to start linking it into Web information sources that live in what is known as the "Linked Open Data Cloud." This is a set of datasets from a number of domains that have partial mappings to other datasets, so that, in essence, developers can mashup the data and then write Web Applications on top of it. In the past year, we've been able to convert a lot of data into Semantic Web formats and to show the power of data mashups, and we've got a lot of really cool things that'll be along soon. In fact, many of our mashups are now linked from the US Data.gov data-sharing website, and there is an area on that site (<http://data.gov/semantic>) which links to our page and features several projects we are working on.

The other project is sort of at the other end of the scale. We've been using the supercomputers available to us at RPI's Computational Center for Nanotechnology³ Innovations to explore scaling the algorithms that power Semantic Web applications to really large datasets. We've been playing with graphs that have many billions of RDF "triples" (essentially the assertions underlying the Semantic Web stuff) and exploring how we can process them in a number of different and interesting ways. There's really only a small number of groups working on this approach, and there's only a couple of US groups in the space, so it is great fun. Turns out we get really nice parallelization on a number of processes, which speaks well to these algorithms eventually moving to multicore machines and to the sorts of backend server farms that power large web applications with millions of users and are the backbone of cloud technologies.

Q. What would you identify as the major challenges you are facing in your work, in the near term and in the long term?

Near term the issue is staying ahead of the commercial world. I mention above that we're starting to play with billions of triples but, for example, the Open Calais project (<http://www.opencalais.com/>) which is just one of many new projects playing with these technologies blogs that they are creating about 750-800 million triples a week⁴! So in that way the Web has of making scale critical, the numbers are growing really big really fast. A second issue is that a lot of the power will come when applications start doing more linking to other applications

³ <http://www.rpi.edu/research/ccni/>

⁴ <http://www.semanticsincorporated.com/2008/10/thomas-tague-comments-on-the-use-of-predicates-in-calais-semantic-proxy.html>

through the Semantic Web layer. Just as the Web really became visible when a lot of "intranets" started opening up and linking to each other, as I mentioned above, the big Web 3.0 applications are still mostly functioning as separate and non-linking apps. Getting people to understand why the linking is so important, and what the network effect gets you, is a major part of my current "evangelism" efforts.

Q. OK, so what does the future hold for the Web?

I think the Web doesn't look all that different, but applications that are similar to the ones you use now will start seeming to have a lot more data available (expect to see graphs, tables and structured information in a lot more places), will seem to have search-like capabilities (such as binto's matching) that are way beyond the current capabilities, and will increasingly be able to exploit the context of your queries (i.e. right now when you search on the name of a restaurant, your search engine doesn't know if you're looking to choose a restaurant, find out more about a particular restaurant, or are in that restaurant looking for other things nearby).

I also think there is another very important thing that will be different, which is that much much more of your access to the Web will be from your mobile device (the thing currently known as a cell phone) and your location and social context will be much more available to applications your willing to make it available to (the way you can now give your iphone permission to use your GPS location in various apps). We're working on a demo in my current lab, in a project headed by my colleague Deborah McGuinness, where a wine-recommender is coupled with a location-aware phone and can access your Facebook information. So your phone could know you are in a particular restaurant, with a particular set of friends, and could use yours and their preferences to pick appropriate wines from the winelist based on what each person is ordering.

So, in essence, I sort of have this vision in my head of us, using our mobile devices, wandering through a Web of information with the ability to somehow find a lot of the right stuff at the right time, based on where we are, what we're doing, and maybe even who we're doing it with. When I moved to RPI, I took a chair called "Tetherless World Professor" - and the more I've come to think about this new vision, the more I like the title. This stuff is still new and exciting, but I look at it this way - I started playing with the Semantic Web back in the 1990s. As a researcher, I'm not content to sit around and exploit Web 3.0 - my job is to help create Web 4.0!k