the semantic web

Summary by Manuel Naujoks (2011)

This summary is based on three papers by Tim Berners-Lee: "Semantic Web Road map" (1998), "The Semantic Web" (2001) and "Creating a Science of the Web" (2006).

As Tim Berners-Lee points out in his paper (Berners-Lee, 1998), the web was designed to store and access information in a decentralized way. Therefore its content is designed to be consumed by humans. This is a major problem when there are attempts to make the web more semantic so that information can be processed by "machines". The way to solve this would be to develop languages for formatting information as Berners-Lee states. Such a language needs to have certain features that Berners-Lee calls "assertion" and "quotation". The latter would be making assertions over other assertions, as he explains. Languages can also be layered on top of each other to specify schemas and even allow logical expressions. Once the information is embedded into a meaningful schema that is understood by machines, a conversion language could be used to associate documents and/or statements with other documents and/or statements. Berners-Lee explains this concept with converting queries for one database into a query for a different and independent database.

According to Berners-Lee basic layer-concepts from bottom to top are assertions, quotations, predicate logic and over that, quantification. Further it would all come down to how to write "the right RDF". He describes the workflow of how machines would process information is starting with identifying data with an URI. That would lead the machine to raw low-level data with a mime-type that is describing what something is. Then the data would be read by an XML parser before it would be read by an RDF parser that produces some kind of graph or logical expression. After that the machine would analyze how the URI is entangled in that graph or formula and dereference other URIs to broaden the graph. That would allow the machine to do some straightforward reasoning. Additionally URIs are seen to also be keys that convey rudimentary trust. In the context of validating an arbitrary proof or a similar input of a user, Berners-Lee writes that there can never be certainty since there are no perfect algorithms for answering arbitrary questions. Constrained rules on the other hand would lead to a chain of assertions that would lead to executable results. These results could proof themself by describing their chain of assertions, their rules and providing references to all the supporting material. He then describes how a language could handle evolution of itself. For example, an implementation that can read a version n should also be able to process a format of version n+1. This kind of backwards compatibility could be achieved by having the version 2 file to provide a reference for the schema of the first version. Further, two independent applications should be able to use themselves sufficiently. That is they both should be able to read and process their schema information and the resulting data.

He also says that the RDF logic level is powerful enough to be used to make these rules that enable what he calls "inference". That is the case because RDF does not rely on heuristics and therefore remains engine agnostic. Even if inference cannot be established successfully (this is what he calls the "annotation problem") a third party index structure could provide connections of all the different schemas. Having this entire infrastructure, it raises the question of how the user provides his input or "query", like Berners-Lee calls it. Since queries are declarative assertions about what is to be returned, RDF should be sufficient at the logical level. However, engines could optimize. A language that describes query engines would allow to search across many other engines and composite search results while extending inference over different engines as well. As reasoning spreads across many independent sources, trust can be a major factor in validating. Public-key cryptography (signatures) can fill that gap and provide some kind of trust for documents. Engines could easily validate information and also find documents that were signed by the same instance. This way, systems can do reasoning about trust systems as well.

Finally Berners-Lee mentions that current search engines may also process RDF objects and that a combination of a reasoning engine and a search engine could help to manage the "combinatorial explosion of possibilities" that arises from the use of these objects. For example a search could start with general indices and then filter out irrelevant information. Engines may not be able to answer arbitrary questions but questions that are real and commonly used by humans would still have a remarkable effect.

In the other paper (Berners-Lee, et al., 2001) Tim Berners-Lee and his fellow authors wrote in a less theoretical style because they give more practical examples. They start with a vision which reminds the reader of a science fiction movie and continue by repeating the main ideas of the "semantic web". After that they introduce the term "agent" as something that gathers data and does reasoning on human behalf. Most information in today's web is inherently designed to be read by humans, not computers. In order to change that, a general structure is needed. Web pages could be written not only with HTML editors but with a special tool for writing semantic web pages. This way, information could be given a well-defined meaning. But there are different kinds of information kinds which introduces a gap between what was primarily designed for human consumption and what was mainly produced for machines. That being said, the semantic web does not use data from human writing or speech.

Then the authors elaborate on the basic ideas of knowledge and machines. They claim that locally and centralized systems that work with artificial intelligence grow uncontrolled and become unmanageable very early. The semantic web approach would not constrain itself and therefore accept that not all questions might be answered by it or that paradoxes still remain unresolved. In today's systems, there are rules and data. Usually data can be transferred but rules stay static. Originally the web needed a central database that maintained an index of its content. Today, search engines provide an almost complete index and make such databases obsolete. As a conclusion, the authors argue that all we need would be a language to express data and rules over data. When the result is processed, added logic should be powerful enough to describe complex attributes but should not let itself being fooled by paradoxes. Since the main field of use is regular questions that humans would usually ask, expected questions are like "x is of the type X" and not like "this sentence is false". While XML would be used to put the data in a structured but meaningless form, RDF would provide the information like what the structure means. An RDF document contains rules that consist of a subject, a verb and an object. Alternatively one could describe it as a thing (subject) that has a property (verb) with a value (object), while each part is represented by a URI. Therefore ambiguity is resolved by having a different URI. The authors state that this would be a natural way of describing data for machines.

While ambiguities can be resolved with URIs, terms with the same meaning but different representations can be resolved by something that the authors introduce as "ontology". Ontology would be a document that specifies the relations between terms in a formal fashion. What the authors introduce as "taxonomy" defines classes of objects and specifies how they are connected. This descriptiveness is necessary since machines to not really understand what any information is. They just understand enough to work with it, so that it makes sense and becomes meaningful for the users. In the context of web pages, each page could have a link to its ontology in order to add machine-understandable meaning to it. According to the authors, a better word for "machine" would be "agent". Agents are being described as programs created by users to do the searching and reasoning. In that process, agents share information with other agents and process data to create dense and high-quality information for the user. To create an agent friendly environment, more and more machine-readable content will be provided in the web. The result an agent comes up with can be explained to the user by showing associated sites and how internal reasoning affected content on these sites. This way, agents have a way to prove their outcome and enable the user to check back on found stuff and/or do corrections. Since an agent can never be entirely certain about what it finds, it should doubt any information by default.

The information provider agents can use should be able to be discovered in a decentralized way. Today's technologies provide this kind of service location based on a special syntax. The semantic web approach should rely solely on the exchange of ontologies. Once an agent retrieved information from any source provider, it might apply some kind of artificial intelligence in order to create chains of values that can be compressed to the dense information that is provided to the user. Like mentioned before, RDF consists of triples of URIs. Given that URIs can point to theoretically everything, agents could also do reasoning about real-life devices or other physical objects. These devices could advertise their functionality and specifications, like they already passively do in their manuals, to be used by agents in this sematic web. Berners-Lee and the other authors also envision the semantic web to improve human knowledge as a whole instead of single independent tasks for egoistically purposes of individuals. Required extensions can easily be deployed since everything is just an URI. As the authors claim, all it takes is a unifying language that allows for linking all the above mentioned concepts into a universal web. That would help humans to do meaningful analysis of any kind and create new tools that provide humans with better life, work and learning aid.

The appended research article (Ossenbruggen, et al., 2002) offers inside in to how ontology layers can be defined and what languages can be used for that. It also elaborates the evolution of hypertext, which is used as a term for data with implicit information, to hypermedia, which is intrinsically more explicitly semantic. The paper, of which Berners-Lee was not an author, also illustrates some open questions in this research field. The often mentioned concept of URIs linking from one thing to another is inherently just a one-way connection. A two-way relationship for linking two resources back and forth would be a better application of connecting information on the web. One way to enable that would be to store all the links in a centralized entity and just include references to them instead of embedding links inline of documents. Another problem, as the authors of this paper argue, would be time. Especially if it is time that identifies information in audio or video format.

In the third paper (Berners-Lee, et al., 2006) Berners-Lee and his coauthors try to raise awareness for a "science of the web" in general. This research would aim at getting a better understanding of what the web is right now, how it is evolving and what its true potential could be. It also includes social values in contrast to the web, like trustworthiness, privacy and respect for "social borders". The main idea would be to find small rules that lead to behavior that can be observed when its application is amplified to a greater scale. Applied to analyzing computer behavior, the result would be to find better algorithms and languages. Once a basic understanding of the web could be gained, web science could help to extend the web into the right direction. Workshops in this field showed that this kind of science can result in a very interesting outcome. As things evolve, web science would also research how protecting intellectual property and similar legal challenges can be made feasible in times of powerful tools like the semantic web. A first solution would be to create the infrastructure in a policy aware fashion, as the authors mention. In summary, web science would be finding new protocols, understanding society using them and find out how that usage could be made highly beneficial for humans.

Works Cited

Berners-Lee, Tim. 1998. *Semantic Web Road map.* 1998.

Berners-Lee, Tim, et al. 2006. *Creating a Science of the Web.* 2006.

Berners-Lee, Tim, Hendler, James and Lassila, Ora. 2001. *The Semantic Web.* 2001.

Ossenbruggen, Jacco van, Hardman, Lynda and Rutledge, Lloyd. 2002. *Hypermedia and the Semantic Web.* 2002.