Rule-based intelligence in the Semantic Web

-or-

"I'll settle for a web that's just not so dumb!"

Dean Allemang
TopQuadrant, Inc.
dallemang@topquadrant.com

Abstract

One could fairly say that the role of rules in the semantic web has been controversial; in the few short years since the first publication of the Semantic Web stack, Rules have sometimes been given a central role, at other times a peripheral role, and sometimes left out completely. Why such variation for an technology with thirty years of background?

The reason for these differences of opinion stem from different goals for the inclusion of rules in the Semantic Web stack. At one extreme are the Description Logicians who see no need for a general-purpose programming language in the semantic web stack. At the other extreme are those who want to build a web infrastructure with the capacity for emergent intelligence.

Our experience with deploying semantic web solutions using OWL alongside rules suggests a moderate middle path; we don't need or even want our web infrastructure to exhibit intelligence; that's what our applications are for. We just need a consistent and coherent web of information to work from. Simply put, we just want our web infrastructure not to be so dumb.

Armed with this understanding, we can see the role of rules in the Semantic Web in a different light, and see a clear role of rules in the semantic web stack.

What do we want from the web?

The document web that we enjoy today has allowed individuals to expand their knowledge in ways unimagined even 20 years ago. Everyone with even a passing interest in movies now has the understanding and knowledge of any genre of movies that was available only to the most dedicated film buff years ago. The same is true in music, and in art, and in every human cultural endeavor. The advances in technical fields is even more impressive; scientists and engineers collaborate on a global basis today in a way that has never before been possible.

When we think about what a Semantic Web might add to this, what could we possibly want? What could possibly go beyond providing access to a global repository of all human knowledge, spanning culture, religion, literature, technology, and indeed any area of human endeavor?

The word "Semantic" refers to "meaning" – a semantic web is a web of meaning. A web that knows what the entities on the web mean, and can make use of that knowledge.

Why do we think that the web would be improved, if it understood the meaning of its contents? Doesn't it understand it now? Google is very good at correcting typing mistakes, figuring out what I "meant" when I mis-typed a query. What more could we want from an integrated web of meaning?

There is a wide range of possible answers to this question, some conservative, some quite revolutionary. Some of them have very subtle ramifications in how the web can develop and how semantics should be supported by the web infrastructure. In particular, the role of rules in the semantic web is strongly influenced by expectations of what a semantic web should do and what it is for.

Sometimes it is said that the semantic web will make data become "smart". What would it mean, for data to be smart? In its most extreme form, smart data means that the web of information becomes so richly interconnected that it can support the Kurzweil Singularity [1] [2]. The Singularity is the point at which a network of information spontaneously becomes sentient and intelligent. The increase in

intimacy of connection on the Semantic Web could be seen as a contributor to this phenomenon.

Another understanding of the idea of "smart data" comes from the Google Mystique. This stems from the observation that Google reliably brings up the most relevant web page for any input query, if not as its first hit, then certainly in the top five. It even corrects the query if something is spelled wrong! Given the fact that Google accomplishes this, it is clear that Google understands a lot about the content of the web already, and it understands a lot about my queries. Of course, the pesky fact that it is quite an art to manipulate Google to behave in this way does little to squelch this expectation.

In a more technical situation, say an engineering company with hundreds of thousands or even millions of design specs, maintenance docs, user manuals, change notes and analyses, surely we can move on just a bit, and have the machine understand our business. Once it does this, it can watch over the shoulder of a designer, and suggest changes to what they are doing, or even propose new designs! But in any case, it should be able to automatically read and understand millions of documents, and decide when they are relevant for any particular engineer.

Still another interpretation of a semantic web is based on the idea of a mash-up. We have information from many sources about how genes relate to enzymes relate to pathways etc., but we can't see how they all connect. A simple matter of giving things global names allows a semantic mash-up to display all the information in a single place.

These different visions of the goal of a semantic web have profound impact on the role of rules. In the singularity, rules will participate in the encoding of sentient cognition, as the network "wakes up". In the example of document understanding, rules play at best a passive role, in the case in which they are used to encode the relationships between documents and their content. In the mash-up example, rules can be completely superfluous.

In this talk, I propose that a useful but modest goal for the semantic web is to create an environment where a variety of stakeholders can create knowledge content, and share it in a distributed web of data. That one enabling functionality for this capability is that it be possible to re-interpret data dynamically and post-hoc, that is, after it has been published. The data isn't drawing any profound conclusions (and certainly isn't conscious), but it does get around a number of the situations that make us feel that data is being "dumb".

In short, the useful semantic web is not a web of smart data, but rather a web of data that isn't so dumb.

Smart data and dumb data

Suppose you are investigating our solar system, and hit a comprehensive web site about objects in the system; stars (well, there's just one of those), planets, moons, asteroids and comets are all described there. Each object has its own web page, with photos and essential information (mass, albedo, distance from the sun, shape, size, what object it revolves around, period of rotation, period of revolution, etc.) . At the head of the page is the object category, e.g., planet, moon, asteroid or comet. Another page includes interesting lists of objects; the moons of Jupiter, the named objects in the asteroid belt, the planets that revolve around the sun. This last page has the nine familiar planets, each linked to its own data page. You recently read in the newspaper that the International Astronomical Union (IAU) has decided that the object known as Pluto will no longer be considered a planet! Instead, it will be a "dwarf" planet. You rush to the Pluto page, and see that indeed, the update has been made - Pluto is listed as a dwarf planet. But when you go back to the "Solar Planets" page, you still see nine planets - Pluto is still there. "That's dumb," you think to yourself, "that they didn't update the web pages consistently!"

Data doesn't have to be intelligent to avoid this sort of thing; in fact, well-designed database-backed web sites have very effective techniques for making sure that this sort of thing doesn't happen. But how can this work, if the data comes from multiple source, and is not maintained in one place, or by one agency? When this happens, then it is more likely that simple mis-matches of this sort can occur, and will result in fragile, unreliable, dumb behavior.

Dumb data vs. Smart applications?

The web is full of intelligent applications, with new innovations coming every day. Ideas that once seemed futuristic are now commonplace; search engines make matches that seem deep and intuitive; commerce sites make smart recommendations personalized in uncanny ways to your own purchasing patterns. Mapping sites include detailed information about world geography, and can plan routes and measure distances. The sky is the limit for the technologies a web site can draw upon; from Bayesian networks to neural networks, fuzzy logic to expert systems,

heuristics to genetic algorithms. A rules language for such an application has some very hefty requirements indeed; it has to be able to support the level of intelligent behavior required by the system. This could include complex uncertainty calculations, ruleset modularity, priorities and dependencies.

But what is the role of the web infrastructure in making these applications "smart"? It isn't practical, or even possible, for the web infrastructure to provide specific support for all, or even any, of these technologies. When we speak of a rule language for the semantic web, are not speaking of enabling intelligent applications – rules support for these intelligent applications is pretty much open ended – we are talking about the infrastructural requirements. What is needed from a rule in the web infrastructure?

Even the most insightful and intelligent application is only as smart as the data that is available to it. Inconsistent or contradictory data will still result in confusing, disconnected, "dumb" results. What can the web infrastructure provide to improve the consistency and availability of web data? In particular, what is the role of a rule in helping make our data stop being dumb?

I propose that it is not the job of the web infrastructure to make data smart so that intelligent applications can be implemented in the web infrastructure. But neither is it sufficient for the web infrastructure to allow data to be dumb. The responsibility of the web infrastructure is to keep the data from being so dumb that it cripples the effectiveness of intelligent applications.

Rules and Data

It is a clear role, in that rules can express the relationships between data items, even data items from disparate sources. But it is also a modest role, in that there is no need for rules to synthesize novel information from its source data. In fact, the role played by rules can be quite modest indeed.

In our work helping customers make their data less dumb, we have found that very simple rules can express the relationships between data that are needed to keep it from being dumb.

The key to making a semantic web application succeed does not lie in teasing out meaning from documents or other artifacts where the meaning is obscured by its context or by natural language conventions. The key thing to understand about the web, either the document web or the semantic web, is that it allows people to express themselves. It works because there are people with a vested interest in making their ideas known. In the world wide web, this has resulted in an explosion of personal web pages and contributions to about.com and wikipedia; in an intranet, it allows knowledge workers to advertise the results of their work, and to allow them a feeling of accomplishment as their message is received and used.

The role of rules in this scenario is simply a linking role; a way to make sure that information from one place makes it to another. If A links to B, and B links to C, then maybe we can specify a rule to tell us the circumstances in which A should link to C, and how it should link.

Conclusions

The main ramification of the idea of "data shouldn't be dumb" on rules languages is that the language has to be simple, and able to cover simple linkage situations. More complex applications of rules should be reserved for the intelligent applications that will populate the web. This isn't to say that research and standardization on more powerful and flexible rule languages isn't valuable or shouldn't be done, it just isn't necessary or even desirable for the rule-based infrastructure of the seamantic web. Keeping these things separate – the infrastructure from the residents of the web – lets us have a much simpler rule language as part of the web infrastructure.

[1] Raymond Kurzweil and Max More, "Max More and Ray Kurzweil on the Singularity", *KurzweilAI.net*http://www.kurzweilai.net/meme/frame.html?main=/articles/art0408.html?.

[2] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology*, Penguin Group, New York, NY, 2005.