Editor: Charles Petrie • petrie@stanford.edu

Rethinking the Semantic Web, Part 2

Rob McCool • *Yahoo*

The Semantic Web[1] has failed to produce communities, quantity, or quality. Very little public information is presently available in the Resource Description Format (RDF) or Web Ontology Language (OWL) data formats, and Semantic Web services are few and far between. This leaves the existing Web with very little semantic structure.

In the last issue (November/December 2005, pp. 86–87), I examined the problems facing the Semantic Web effort. Here, I propose a solution to the problem of introducing widespread semantics to the World Wide Web. My proposal is to do for the Semantic Web what Tim Berners-Lee (www.w3.org/People/Berners-Lee) did for Project Xanadu, the original hypertext project (www.xanadu.com), and Standardized General Markup Language (www.w3.org/MarkUp/SGML/).

Lessons in Simplicity

When Berners-Lee developed the Web, he took the salient ideas of hypertext and SGML syntax and removed complexities such as backward hyperlinks. At the time, many criticized their absence from HTML because, without them, pages can simply vanish and links can break. But the need to control both the linking and linked pages is a burden to authoring, sharing, and copying.

Similarly, early forms of HTML paid no regard to SGML document-type definitions (DTDs). Berners-Lee simply ignored these difficult to create and understand declarations of how markup tags are used. Semantic Web proponents should take a lesson from this and more recent Web-based successes.

Two recent Web communities are worth examining in this regard. The Flickr photo-sharing site (www.flickr.com) has developed a community of people who "tag" their images with keywords for high-quality image retrieval. The system is almost suicidally simple — there's no notion of synonyms or disambiguation — but simplicity encourages participation. From an academic standpoint, the lack of semantics in tags implies that the system should collapse into incoherence with even moderate participation. Yet, it is one of the most successful Web communities today.

For a Web community with simple, easy-to-use authoring tools that support synonyms, disambiguation, and categories, we can look to Wikipedia, the Internet encyclopedia (http://en.wikipedia.org). It's based on wiki technology, which lets people write HTML in plaintext, as if writing an email — automatically generating bullet points, for example, by starting a line with an asterisk — and supports a very simple set of semantics. Wikipedia calls synonyms *redirect pages*, and disambiguation is explicitly handled via special pages.

Wikipedia's major limitation is that it's a "shadow" web. It will always stand away from its source material and context and will never consistently provide references to its information sources. It would be far better if the information in Wikipedia were spread across the Web, so that users could evaluate authorship and context.

A NEW Approach

The Semantic Web formats must be simplified. Relations and detailed classifications are significant barriers to adoption because they create huge translation and maintenance burdens. Moreover, authors can no longer embed information in language text because it must be formatted as *triples*. A triple is a subject entity, a property, and a value that's either a string literal or another entity.

Removing classes, relations, and triples from Semantic Web formats would give us what I call a *named-entity web* (NEW). This would represent a major leap over today's Web semantics by removing barriers to adoption and enabling applications that today's meager Semantic Web deployment can't.

Adding parameters to existing tags would let us embed named-entity information directly as HTML markup, thus allowing participation by language processors, which have excellent entity recognizers but poor relation recognizers. Natural language processors, such as IBM's Unstructured Information Management Architecture framework (UIMA ; www.research.ibm.com/UIMA/), would be able to read the actual text, as they can do today with Web sites, and use hints from the NEW tags to guide their analysis. They would then be able to embed their output directly into the page's tags. Their participation in this web would thus be very direct. Rather than a shadow web maintained separately from natural language, the Semantic Web would

continued on p. 93

 Published by the IEEE Computer Society 1089-7801/06/$20.00 © 2006 IEEE

continued from p. 96

simply be an extension of the existing Web. The authoring burden would be very low, and basic name and category semantics, trust mechanisms used in the blog world, and sorting strategies used by search engines could all contribute their techniques toward solving the problems of relevance and "spam" detection and prevention.

Each NEW page would be HTML with a `<meta>` tag in its header indicating what the page was about. The NEW would be restricted to pages that were about a single particular entity — for example, a person's home page or a place such as China. The page would provide the entity's type from a simple set, which might initially include "`person`," "`organization`," "`event`," and "`thing`." To embed more than one entity in a single page, we'd use named anchors.

A reasonable response to this system would be to say that I am naively reconstructing ontologies, and that this system will eventually become ontological anyway. My response is that I have a full understanding of and experience with ontologies, and just as HTML didn't need backward hyperlinks for consistency, and just as the chaos of tagging systems like Flickr are counterintuitively effective, NEW doesn't need the consistency and formalism that ontologies work so hard to ensure. In fact, it's vital that the hard work be removed in favor of participation.

An entity page could declare canonical and alternative names. The page URL would be the symbol that represented the entity, but the page would declare human-readable names via `<title>` or `<a>` tags (for visible names) and the `<meta>` tag (for invisible ones). Multiple names would be used for nicknames, abbreviations, and other specialized synonyms.

Given that visual disambiguation is critical for usability, a page could declare images of the entity it defined — visibly, using the `<img>` tag, or invisibly via the `<link>` tag. The page might mark one of its paragraphs as a description of the entity through an extra parameter for text container tags. In this way, a simple business-card style summary of an entity could be generated, and a set of these cards could be presented to users to choose which entity they're interested in.

NEW pages could use the `<link>` tag to declare themselves to be about the same entity as another page. When changing workplaces or ISPs, for example, I could point to my old page and prevent broken links. A shopping site with information about a particular product could declare its entity to be the same as one defined on the manufacturer's product page, thus enabling better search.

An HTML page that defines an entity includes *identifying references* — unnamed relations to other entities that define and disambiguate the entity from others with the same name. For people, these might include workplaces, family members, or friends, whereas companies might refer to locations, executives, or products, and cities might refer to their nations or famous landmarks. For usability, identifying references should be made with `<a>` tags, so that authors can maintain them along with the page text, but they can also be made invisibly using the `<link>` tag.

A page can vouch for another page's integrity with the `<a>` tag, or invisibly with the `<link>` tag. Any entity-defining HTML page that makes a hyperlink to another entity-defining HTML page boosts the latter's authenticity by applying two of the Web's most successful quality-enhancement systems — the PageRank algorithm (www.google.com/corporate/tech.html) and the Technorati blog-aggregation service (www.technorati.com) — to the Semantic Web.

Without such a radical simplification, the Semantic Web will continue to see limited participation and few compelling applications.

Real Applications

Maintaining entity information in HTML allows the Semantic Web to benefit from trust mechanisms developed for the existing Web. Until now, most trust models developed for the Semantic Web have required the deployment of a shadow network of users and trust, much as it relies on a shadow web.

One criticism of the Semantic Web is that the example applications are very contrived. We either get tired examples involving travel, appointments, or book sales or "truly revolutionary" examples that are akin to visions in Bill Gates' book[2] — they're just not going to happen, at least not until the representation problem is solved. NEW applications aspire to similar goals as the less ambitious Semantic Web examples, but would have a better chance of being deployed due to the simplistic, potentially inconsistent structure of the system. Community participation will again be the yardstick by which the success of the system is judged.

The applications for NEW are about people, places, and products. If a large number of people declare entities for themselves, and if search engines pay attention to the entity tags, then people with common names will be easier to find. Named-entity disambiguation and synonyms will thus become part of the fabric of the Web.

The friend-of-a-friend project (www.

continued on p. 94

An Example for the Named-Entity Web

To see how NEW might play out in practice, suppose I made the following my home page at Stanford:

```
<html>
<title>Rob McCool's Home Page</title>
<meta name="entity-name" content="Robert McCool">
<meta name="entity-type" content="person">

<img align=right rel="entity-image" src="robm.jpg">
<p><a rel="entity-name-preferred">Rob McCool</a> is
   an <a    rel="entity-type">AI researcher</a> at <a
    rel="entity-ident"    href="http://www.stanford.
   edu/">Stanford University</a>'s <a rel="entity-
   ident" href="http://ksl.stanford.edu/">Knowledge
   Systems Laboratory</a>.</p>
</html>
```

Suppose the Stanford Knowledge Systems Laboratory (KSL) then added some tags to its home page.

```
<html>
<title>Knowledge Systems Laboratory Home Page</title>
<meta name="entity-name-preferred"
  content="Stanford University Knowledge Systems
   Laboratory">
<meta name="entity-name" content="Knowledge Systems
   Laboratory">
<meta name="entity-type" content="organization">
<p><a rel="entity-name">KSL</a> conducts research in
   the areas of <a rel="entity-ident">knowledge
   representation</a> and <a rel="entity-ident">
   automated reasoning</a> in the <a rel="entity-
   ident" href="http://ai.stanford.edu/">Artificial
   Intelligence Laboratory</a> of the <a rel="entity-
   ident" href="http://cs.stanford.edu/">Department of
   Computer Science</a> at <a rel="entity-ident" href=
   "http://www.stanford.edu/">Stanford University</a>.
</p>

<p>KSL people:</p>
<ul>
<li><a rel=entity-ident href=robm.html>Rob McCool
</a></li>
...
</ul>
</html>
```

With these additional tags, search engines and other information aggregators could gather information that would make me locatable by either my nickname, "Rob McCool," or my full name, "Robert McCool." With sufficient participation, users could find "people doing AI research in California" via search engines. With the knowledge that Stanford is in California (available in the major Local/Maps offerings these days), and the information in these tags, we could get answers to semantic queries that are currently impossible.

continued from p. 93

foaf-project.org) also becomes integral to the Web, rather than being another shadow web. Running on blog entries, news stories, and other documents, language processors would classify references to named entities and use an index to provide documents about entities, not just keywords. Search engine queries could combine words and URLs so that nobody would be punished with fewer results for typing "`Ed Feigenbaum`" instead of "`Edward Feigenbaum`."

For products, manufacturers could define product entities, to which online retailers could refer in their pages. Shopping search would become more accurate as manufacturers' product pages became their symbols. Manufacturers could define canonical names for their products and shortcuts, such as series designations, to help users find products more easily. With NEW, users could specify particular products they were seeking and thus avoid high-ranked pages that happened to have the same words as those contained in the product names.

NEW also has mapping and location applications. Today, local searches are seeded with phone book listings and augmented with address crawlers. Creating a specific link between a business and the city it's located in would provide users with better results. NEW-aware mapping services could use cities' official home pages as canonical URLs to provide businesses the option of using their location as a jumping point to a world of information. Combining longitudes, latitudes, and city names in current mapping services with symbolic representations of those places would open paths along the Web for developers to use in creating place awareness. Taken further, users strolling a city with a GPS device could access pages that could lead them through the location like a tour guide. Entities could thus help them find higher-level information than any map can provide.

NEW would make use of existing Web technologies and provide direct benefits with a far lower participation cost than current semantic technologies require. Without such a radical simplification, the Semantic Web will continue to see limited participation and few compelling applications.

Given that the lowest common denominator, backed by the most cash, tends to win, another possibility is that

Google Base (http://base.google.com/) will be the Semantic Web in the end. Not only can they build it themselves, but they own the 'shadow web' and can exploit it in ways that others can't. The tool combines simple tagging with a tool for creating dynamic schemas, and for entering database entries. It even includes a simple class system.

If I'm wrong, either Google Base or the Semantic Web will dominate over something HTML-integrated like NEW. The wildcard here is that Google can bring all of its technology to bear on the shadow web problem. Having a shadow web isn't a problem if you own it, monetize it, allow others to monetize it, and can solve the spam problem because you own the human Web-based spam-remedy systems (as opposed to the Semantic Web, in which someone must first gather the data and then solve the spam problem). As with the Semantic Web, community participation will determine whether Google Base achieves its ambitions.

Acknowledgments

The opinions in this column are mine. They do not necessarily reflect those of my employer.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web: A New Form of Web Content that Is Meaningful to Computers Will Unleash a Revolution of New Possibilities," *Scientific Am.*, May 2001, pp. 28–37; www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
2. B. Gates, N. Myhrvold, and P. Rinearson, *The Road Ahead*, Penguin, 1996.

Rob McCool is a software developer and architect at Yahoo. He worked on the Alpiri Project (TAP) on Semantic Web technologies at Stanford University from 2002 until earlier in 2005. He also authored the US National Center for Supercomputing Applications (NCSA) httpd Web server, which became the Apache Web server. McCool has a BS in computer science from the University of Illinois at Urbana-Champaign. Contact him at robm@robm.com.