Structure Paves the Way to the Semantic Web

by Mike Bergman - Thursday, May 03, 2007

http://www.mkbergman.com/357/structure-paves-the-way-to-the-semantic-web/

[Click on image for full-size pop-up]

Colorado Interstate Construction - 1970
Courtesy National Archives

NOTE: I was pleased when Jim Hendler asked me to pen some thoughts on the semantic Web (as a vehement moderate, I always use the mixed case). I think both of us hoped that my background combining Internet business with Web science might bring a pragmatic perspective to the subject. The material below is to appear as a guest editorial in IEEE Intelligent Systems in the May/June issue. I thank it for allowing early releases by authors, and to Linda World, its super senior editor, for cleaning up my language. There are some other differences due to length considerations. ***MKB***

For a dozen years, my career has been centered on Internet search, dynamic content and the deep Web. For the past few years, I have been somewhat obsessed by two topics. The first topic, a conviction really, is that implicit structure needs to be extracted from Web content to enable it to be disambiguated, organized, shared and re-purposed. The second topic, more an open question as a former academic married to a professor, is what might replace editorial selections and peer review to establish the authoritativeness of content. These topics naturally steer one to the semantic Web.

A Millennial Perspective

The semantic Web, by whatever name it comes to be called, is an inevitability. History tells us that as information content grows, so do the mechanisms for organizing and managing it. Over human history, innovations such as writing systems, alphabetization, pagination, tables of contents, indexes, concordances, reference look-ups, classification systems, tables, figures, and statistics have emerged in

parallel with content growth.

When the Lycos search engine, one of the first profitable Internet ventures, was publicly released in 1994, it indexed a mere 54,000 pages [1]. When Google wowed us with its page-ranking algorithm in 1998, it soon replaced my then favorite search engine, AltaVista. Now, tens of billions of indexed documents later, I often find Google's results to be overwhelming dross -- unfortunately true again for all of the major search engines. Faceted browsing, vertical search, and Web 2.0's tagging and folksonomies demonstrate humanity's natural penchant to fight this entropy, efforts that will next continue with the semantic Web and then mechanisms unforeseen to manage the chaos of burgeoning content.

An awful lot of hot air has been expelled over the false dichotomy of whether the semantic Web will fail or is on the verge of nirvana. Arguments extend from the epistemological versus ontological (classically defined) to Web 3.0 versus SemWeb or Web services (WS*) versus REST (Representational State Transfer). My RSS feed reader points to at least one such dust up every week.

Some set the difficulties of resolving semantic heterogeneities as absolutes, leading to an illogical and false rejection of semantic Web objectives. In contrast, some advocates set equally divisive arguments for semantic Web purity by insisting on formal ontologies and descriptive logics. Meanwhile, studied leaks about "stealth" semantic Web ventures mean you should grab your wallet while simultaneously shaking your head.

A Decades-Long Perspective

My mental image of the semantic Web is a road from here to some achievable destination--say, Detroit. Parts of the road are well paved; indeed, portions are already superhighways with controlled on-ramps and off-ramps. Other portions are two lanes, some with way too many traffic lights and some with dangerous intersections. A few small portions remain unpaved gravel and rough going.

[Click on image for full-size pop-up]

Wreck in Nebraska during the 1919 Transcontinental Motor Convoy
Courtesy National Archives

A lack of perspective makes things appear either too close or too far away. The automobile isn't yet a century old as a mass-produced item. It wasn't until 1919 that the US Army Transcontinental Motor Convoy made the first automobile trip across the United States. The 3,200 mile route roughly followed today's Lincoln Highway, US 30, from Washington, D.C. to San Francisco. The convoy took 62 days and 250 recorded accidents to complete the trip (see figure), half on dirt roads at an average speed of 6 miles per hour. A tank officer on that trip later observed Germany's autobahns during World War II. When he subsequently became President Dwight D. Eisenhower, he proposed and then signed the Interstate Highway Act. That was 50 years ago. Today, the US is crisscrossed with 50,000 miles of interstates, which have completely remade the nation's economy and culture [2].

Today's Perspective

Like the interstate system in its early years, today's semantic Web lets you link together a complete trip, but the going isn't as smooth or as fast as it could be. Nevertheless, making the trip is doable and keeps improving day by day, month by month.

My view of what's required to smooth the road begins with extracting structure and meaningful information according to understandable schema from mostly uncharacterized content. Then we store the now-structured content as RDF triples that can be further managed and manipulated at scale. By necessity, the journey embraces tools and requirements that, individually, might not constitute semantic Web technology as some strictly define it. These tools and requirements are nonetheless integral to reaching the destination. We are well into that journey's first leg, what I and others are calling the *structured Web*.

For the past six months or so I have been researching and assembling as many semantic Web and related tools as I can find [3]. That ***Sweet Tools*** listing now exceeds 500 tools [4] (with its presentation using the nifty lightweight Exhibit publication system from MIT's Simile program [5]). I've come to understand the importance of many ancillary tool sets to the entire semantic Web highway, such as natural language processing and information extraction. I've also found new categories of pragmatic tools that embody semantic Web and data mediation processes but don't label themselves as such.

In its entirety, the ***Sweet Tools*** listing provides a pretty good picture of the semantic Web's state. It's a surprisingly robust picture -- though with some notable potholes -- and includes impressive open source options in all categories. Content publishing, indexing, and retrieval at massive scales are largely solved problems. We also have the infrastructure, languages, and (yes!) standards for tying this content together meaningfully at the data and object levels.

I also think a degree of consensus has emerged on RDF as the canonical data model for semantic information. RDF triple stores are rapidly improving toward industrial strength, and RESTful designs enable massive scalability, as terabyte- and petabyte-scale full-text indexes prove.

Powerful and flexible middleware options, such as those from OpenLink [6], can transform and integrate diverse file formats with a variety of back ends. The World Wide Web Consortium's GRDDL standard [7] and related tools, plus various "RDF-izers" from Massachusetts Institute of Technology and elsewhere

[8], largely provide the conversion infrastructure for getting Web data into that canonical RDF form. Sure, some of these converters are still research-grade, but getting them to operational capabilities at scale now appears trivial.

Things start getting shakier when trying to structure information into a semantic formalism. Controlled vocabularies and ontologies range broadly and remain a contentious area. Publishers and authors perhaps have too many choices: from straight Atom or RSS feeds and feeds with tags to informal folksonomies and then Outline Processor Markup Language [9] or microformats [10]. From there, the formalism increases further to include the standard RDF ontologies such as SIOC (Semantically-Interlinked Online Communities), SKOS (Simple Knowledge Organizing System), DOAP (Description of a Project), and FOAF (Friend of a Friend) [11] and the still greater formalism of OWL's various dialects [12].

If we compare the semantic Web to the US interstate highway system, we're still in the early stages of a journey that will remake our economy and culture.

Many potholes on the road to the semantic Web exist.

One ready task is to transform existing structure to RDF. Another priority is to refine tools to extract structure and meaningful information from uncharacterized content.

Arguing which of these is the theoretical best method is doomed to failure, except possibly in a bounded enterprise environment. We live in the real world, where multiple options will always have their advocates and their applications. All of us should welcome whatever structure we can add to our information base, no matter where it comes from or how it's done. The sooner we can embrace content in any of these formats and convert it into canonical RDF form, we can then move on to needed developments in semantic mediation, some of the roughest road on the journey.

Potholes on the Semantic Highway

Semantic mediation requires appropriate structured content. Many potholes on the road to the semantic Web exist because the content lacks structured markup; others arise because existing structure requires transformation. We need improved ways to address both problems. We also need more intuitive means for applying schema to structure. Some have referred to these issues as "who pays the tax."

Recent experience with social software and collaboration proves that a portion of the Internet user community is willing to tag and characterize content. Furthermore, we can readily leverage that resulting structure, and free riders are welcomed. The real pothole is the lack of easy--even fun--data extractors and "structurizers." But we're tantalizingly close.

Tools such as Solvent and Sifter from MIT's Simile program [13] and Marmite from Carnegie Mellon University [14] are showing the way to match DOM (document object model) inspectors with automated structure extractors. DBpedia, the alpha version of Freebase, and System One now provide large-scale, open Web data sets in RDF [15], including all of Wikipedia. Browser extensions such as Zotero [16] are

showing how to integrate structure management into acceptable user interfaces, as are services such as Zoominfo [17]. Yet we still lack easy means to design the differing structures suitable for a plenitude of destinations.

Amazingly, a compelling road map for how all these pieces could truly fit together is also incomplete. How do we actually get from here to Detroit? Within specific components, architectural understandings are sometimes OK (although documentation is usually awful for open source projects, as most of the current tools are). Until our community better documents that vision, attracting new contributors will be needlessly slower, thus delaying the benefits of network effects.

So, let's create a road map and get on with paving the gaps and filling the potholes. It's not a matter of standards or technology--we have those in abundance. Let's stop the silly squabbles and commit to the journey in earnest. The *structured Web*'s ability to reach *Hyperland* [18], Douglas Adam's prescient 1990 forecast of the semantic Web, now looks to be no further away than Detroit.

---

[1] Chris Sherman, "Happy Birthday, Lycos!," *Search Engine Watch*, August 14, 2002. See http://searchenginewatch.com/showPage.html?page=2160551.

[2] David A. Pfeiffer, "Ike's Interstates at 50: Anniversary of the Highway System Recalls Eisenhower's Role as Catalyst," *Prologue Magazine*, National Archives, Summer 2006, Vol. 38, No. 2. See: http://www.archives.gov/publications/prologue/2006/summer/interstates.html.

[3] The mention of specific tool names is meant to be illustrative and not necessarily a recommendation.

[4] **Sweet Tools** (SemWeb) listing; see http://www.mkbergman.com/?page_id=325 .

[5] See http://simile.mit.edu/exhibit/.

[6] OpenLink Software's Virtuoso and Data Spaces products; see http://www.openlinksw.com/.

[7] W3C's Gleaning Resource Descriptions from Dialects of Languages (GRDDL, pronounced "griddle"). See http://www.w3.org/2004/01/rdxh/spec.

[8] See http://simile.mit.edu/wiki/RDFizers.

[9] Outline Processor Markup Language (OPML); see http://www.opml.org/.

[10] Microformats; see http://microformats.org/.

[11] DOAP (Description of a Project), FOAF (Friend of a Friend), SIOC (Semantically-Interlinked

Online Communities) and SKOS (Simple Knowledge Organizing System)..

[12] W3C's Web Ontology Language (OWL). See http://www.w3.org/TR/owl-features/.

[13] Solvent (http://simile.mit.edu/wiki/Solvent) and Sifter (http://simile.mit.edu/wiki/Sifter) are from MIT's Simile program.

[14] Marmite (http://www.cs.cmu.edu/~jasonh/projects/marmite/) is from Carnegie Mellon University.

[15] DBpedia (http://dbpedia.org/docs/) and Freebase (in alpha, by invitation only at http://www.freebase.com/) are two of the first large-scale open datasets on the Web; Wikipedia has also been converted to RDF by System One (http://labs.systemone.at/wikipedia3).

[16] Zotero is produced by George Mason University's Center for History and New Media; see http://www.zotero.org.

[17] ZoomInfo (http://www.zoominfo.com/) provides online structured search of companies and people, plus broader services to enterprises.

[18] The late Douglas Adams, of *Doctor Who* and *A Hitchhiker's Guide to the Galaxy* fame, produced a TV program for BBC2 presaging the Internet called Hyperland. This 50-min video can be seen in five parts via YouTube at Part 1 of 5, 2 of 5, 3 of 5, 4 of 5 and 5 of 5.
