Seeking a Semantic Web Sweet Spot

by Mike Bergman - Monday, February 21, 2011

http://www.mkbergman.com/946/seeking-a-semantic-web-sweet-spot/

Reference Structures Provide a Third Way

Since the first days of the Web there has been an ideal that its content could extend beyond documents and become a global, interoperating storehouse of data. This ideal has become what is known as the "semantic Web". And within this ideal there has been a tension between two competing world views of how to achieve this vision. At the risk of being simplistic, we can describe these world views as *informal* v *formal*, sometimes expressed as "bottom up" v "top down" [1,2].

The *informal* view emphasizes freeform and diversity, using more open tagging and a bottoms-up approach to structuring data [3]. This group is not anarchic, but it does support the idea of open data, open standards and open contributions. This group tends to be oriented to RDF and is (paradoxically) often not very open to non-RDF structured data forms (as, for example, microdata or microformats). Social networks and linked data are quite central to this group. RDFa, tagging, user-generated content and folksonomies are also key emphases and contributions.

The *formal* view tends to support more strongly the idea of shared vocabularies with more formalized semantics and design. This group uses and contributes to open standards, but is also open to proprietary data and structures. Enterprises and industry groups with standard controlled vocabularies and interchange languages (often XML-based) more typically reside in this group. OWL and rules languages are more often typically the basis for this group's formalisms. The formal view also tends to split further into two camps: one that is more top down and engineering oriented, with typically a more *closed world approach* to schema and ontology development [4]; and a second that is more adaptive and incremental and relies on an *open world approach* [5].

Again, at the risk of being simplistic, the informal group tends to view many OWL and structured vocabularies, especially those that are large or complex, as over engineered, constraining or limiting

freedom. This group often correctly points to the delays and lack of adoption associated with more formal efforts. The informal group rarely speaks of ontologies, preferring to use the term of vocabularies. In contrast, the formal group tends to view bottoms-up efforts as chaotic, poorly structured and too heterogeneous to allow machine reasoning or interoperability. Some in the formal group sometimes advocate certification or prescribed training programs for ontologists.

Readers of this blog and customers of Structured Dynamics know that we more often focus on the *formal* world view and more specifically from an open world perspective. But, like human tribes or different cultures, there is no one true or correct way. Peaceful coexistence resides in the understanding of the importance and strength of different world views.

Shared communication is the way in which we, as humans, learn to understand and bridge cultural and tribal differences. These very same bases can be used to bridge the differences of world views for the semantic Web. Shared concepts and a way to communicate them (via a common language) -- what I call *reference structures* [6] -- are one potential "sweet spot" for bridging these views of the semantic Web [7] .

Referring to Referents as Reference

According to Merriam Webster and Wikipedia, a *reference* is the intentional use of one thing, a point of reference or reference state, to indicate something else. When reference is intended, what the reference points to is called the *referent*. References are indicated by sounds (like onomatopoeia), pictures (like roadsigns), text (like bibliographies), indexes (by number) and objects (a wedding ring), but many other methods can be used intentionally as references. In language and libraries, references may include dictionaries, thesauri and encyclopedias. In computer science, references may include pointers, addresses or linked lists. In semantics, reference is generally construed as the relationships between nouns or pronouns and objects that are named by them.

The Building Blocks of Language

Structures, or syntax, enable multiple referents to be combined into more complex and meaningful (interpretable) systems. Vocabularies refer to the set of tokens or words available to act as referents in these structures. Controlled vocabularies attempt to limit and precisely define these tokens as a means of reducing ambiguity and error. Larger vocabularies increase richness and nuance of meaning for the tokens. Combined, syntax, grammar and vocabularies are the building blocks for constructing understandable human languages.

Many researchers believe that language is an inherent capability of humans, especially including children. Language acquisition is expressly understood to be the combined acquisition of syntax, vocabulary and phonetics (for spoken language). Language development occurs via use and repetition, in a social setting where errors are corrected and communication is a constant. Via communication and interaction we learn and discover nuance and differences, and acquire more complex understandings of syntax structures and vocabulary. The contact sport of communication is itself a prime source for acquiring the ability to communicate. Without the structure (syntax) and vocabulary acquired through this process, our language utterances are mere babblings.

Pidgin languages emerge when two parties try to communicate, but do not share the same language. Pidgin languages result in much simplified vocabularies and structure, which lead to frequent miscommunication. Small vocabularies and limited structure share many of these same limitations.

Communicating in an Evanescent Environment

Information theory going back to Shannon defined that the "fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point" [8]. This assertion applies to all forms of communication, from the electronic to animal and human language and speech.

Every living language is undergoing constant growth and change. Current events and culture are one driver of new vocabulary and constructs. We all know the apocryphal observation that northern peoples have many more words for snow, for example. Jargon emerges because specific activities, professions, groups, or events (including technical change) often have their own ideas to communicate. Slang is local or cultural usage that provides context and communication, often outside of "formal" or accepted vocabularies. These sources of environmental and other changes cause living languages to be constantly changing in terms of vocabulary and (also, sometimes) structure.

Natural languages become rich in meaning and names for entities to describe and discern things, from plants to people. When richness is embedded in structure, contexts can emerge that greatly aid removing ambiguity ("disambiguating"). Contexts enable us to discern polysemous concepts (such as bank for river, money institution or pool shot) or similarly named entities (such as whether Jimmy Johnson is a race car driver, football coach, or a local plumber). As with vocabulary growth, contexts sometimes change in meaning and interpretation over time. It is likely the Gay '90s would not be used again to describe a cultural decade (1890s) in American history.

All this affirms what all of us know about human languages: they are dynamic and changing. Adaptable (living) languages require an openness to changing vocabulary and changing structure. The most dynamic languages also tend to be the most open to the coining of new terminology; English, for example, is estimated to have 25,000 new words coined each year [9].

The Semantic Web as a Human Language

One could argue that similar constructs must be present within the semantic Web to enable either machine or human understanding. At first blush this may sound a bit surprising: Isn't one premise of the semantic Web machine-to-machine communications with "artificial intelligence" acting on our behalf in the background? Well, hmmm, OK, let's probe that thought.

Recall there are different visions about what constitutes the semantic Web. In the most machine-oriented version, the machines are posited to replace some of what we already do and anticipate what we already want. Like Watson on Jeopardy, machines still need to know that Toronto is not an American city [10]. So, even with its most extreme interpretation -- and one that is more extreme than my own view of the near-term semantic Web -- machine-based communication still has these imperatives:

- Humans, too, interact with data and need to understand it

- Much of the data to be understood and managed is based on human text (unstructured), and needs to be adequately captured and represented
- There is no basis to think that machine languages can be any simpler in representing the world than human languages.

These points suggest that machine languages, even in the most extreme machine-to-machine sense, still need to have a considerable capability akin to human languages. Of course, computer programming languages and data exchange languages as artificial languages need not read like a novel. In fact, most artificial languages have more constraints and structure limitations than human languages. They need to be read by machines with fixed instruction sets (that is, they tend to have fewer exceptions and heuristics).

But, even with software or data, people write and interact with these languages, and human readability is a key desirable aspect for modern artificial languages [11]. Further, there are some parts of software or data that also get expressed as labels in user interfaces or for other human factors. The admonition to Web page developers to "view source" is a frequent one. Any communication that is text based -- as are all HTTP communications on the Web, including the semantic Web -- has this readability component.

Though the form (structure) and vocabulary (tokens) of languages geared to machine use and understanding most certainly differ from that used by humans, that does not mean that the imperatives for reference and structure are excused. It seems evident that small vocabularies, differing vocabularies and small and incompatible structures have the same limiting effect on communications within the semantic Web as they do for human languages.

Yet, that being said, correcting today's relative absence of reference and structure on the nascent semantic Web should not then mean an overreaction to a solution based on a single global structure. This is a false choice and a false dichotomy, belied by the continued diversity of human languages [12]. In fact, the best analog for an effective semantic Web might be human languages with their vocabularies, references and structures. Here is where we may find the clues for how we might improve the communications (interoperability) of the semantic Web.

A Call for Vehement Moderation

Freeform tagging and informal approaches are quick and adaptive. But, they lack context, coherence and a basis for interoperability. Highly engineered ontologies capture nuance and sophistication. But, they are difficult and expensive to create, lack adoption and can prove brittle. Neither of these polar opposites is "correct" and each has its uses and importance. Strident advocacy of either extreme alone is shortsighted and unsuited to today's realities. There is not an ineluctable choice between freedom and formalism.

An inherently open and changing world with massive growth of information volumes demands a third way. Reference structures and vocabularies sufficient to *guide* (but not *constrain*) coherent communications are needed. Structure and vocabulary in an open and adaptable language can provide the communication medium. Depending on task, this language can be informal (RDF or data struct forms convertible to RDF) or formal (OWL). The connecting glue is provided by the reference vocabularies and structures that bound that adaptable language. This is the missing "sweet spot" for the semantic Web.

Just like human languages, these reference structures must be adaptable ones that can accommodate new learning, new ideas and new terminology. Yet, they must also have sufficient internal consistency and structure to enable their role as *referents*. And, they need to have a richness of vocabulary (with defined references) sufficient to capture the domain at hand. Otherwise, we end up with pidgin communications.

We can thus see a pattern emerging where *informal* approaches are used for tagging and simple datasets; more *formal* approaches are used for bounded domains and the need for precise semantics; and *reference structures* are used when we want to get multiple, disparate sources to communicate and interoperate. So long as these reference structures are coherent and designed for vocabulary expansion and accommodation for synonyms and other means for terminology mapping, they can adapt to changing knowledge and demands.

For too long there has been a misunderstanding and mischaracterization of anything that smacks of structure and referenceability as an attempt to limit diversity, impose control, or suggest some form of "One Ring to rule them all" organization of the semantic Web. Maybe that was true of other suggestions in the past, but it is far from the enabling role of reference structures advocated herein. This reaction to structure has something of the feeling of school children adverse to their writing lessons taking over the classroom and then saying No! to more lessons. Rather than Lord of the Rings we get Lord of the Flies.

To try to overcome this misunderstanding -- and to embrace the idea of language and communication for the semantic Web -- I and others have tried in the past to find various analogies or imagery to describe the roles of these reference structures. (Again, all of those vagaries of human language and communication!). Analogies for these reference structures have included [13]:

- *backbones*, to signal their importance as dependable structures upon which we can put "meat on the bones"
- *scaffoldings*, to emphasize their openness and infrastructural role
- *roadmaps*, as orienting and navigational frameworks for information
- *docking ports*, as connection points for diverse datasets on the Web
- *forest paths*, to signal common traversals but with much to discover once we step off the paths
- *infoclines*, to represent the information interface between different world views,
- and others.

What this post has argued is the analogy of reference structures to human language and communication. In this role, reference structures should be seen as facilitating and enabling. This is hardly a vision of constraints and control. The ability to articulate positions and ideas in fact leads to more diversity and freedom, not less.

To be sure, there is extra work in using and applying reference structures. Every child comes to know there is work in learning languages and becoming articulate in them. But, as adults, we also come to learn from experience the frustration that individuals with speech or learning impairments have when trying to communicate. Knowing these things, why do we not see the same imperatives for the semantic Web? We can only get beyond incoherent babblings by making the commitment to learn and master rich languages grounded in appropriate reference structures. We are not compelled to be inchoate; nor are our machines.

Yet, because of this extra work, it is also important that we develop and put in place semi-automatic [14]

ways to tag and provide linkages to such reference structures. We have the tools and information extraction techniques available that will allow us to reference and add structure to our content in quick and easy ways. Now is the time to get on with it, and stop babbling about how structure and reference vocabularies may limit our freedoms.

This is the first of a two-part series on the importance and choice of *reference structures* (Part I) and *gold standards* (Part II) on the semantic Web.

---

[1] This is reflected well in a presentation from the NSF Workshop on DB & IS Research for Semantic Web and Enterprises, April 3, 2002, entitled "The "Emergent, Semantic Web: Top Down Design or Bottom Up Consensus?". This report defines top down as design and committee-driven; bottom up is more decentralized and based on social processes. Also, see Ralf Klischewski, 2003. "Top Down or Bottom Up? How to Establish a Common Ground for Semantic Interoperability within e-Government Communities," pp. 17-26, in R. Traunmüller and M Palmirani, eds., *E-Government: Modelling Norms and Concepts as Key Issues: Proceedings of 1st International Workshop on E-Government* at ICAIL 2003, Bologna, Italy. Also, see David Weinberger, 2006. "The Case for Two Semantic Webs," *KM World*, May 26, 2006; see http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=15809.

[2] For a discussion about formalisms and the nature of the Web, see this early report by F.M. Shipman III and C.C. Marshall, 1994. "Formality Considered Harmful: Experiences, Emerging Themes, and Directions," *Xerox PARC Technical Report ISTL-CSA-94-08-02*, 1994; see http://www.csdl.tamu.edu/~shipman/formality-paper/harmful.html.

[3] Others have posited contrasting styles, most often as "top down" *v.* "bottom up." However, in one interpretation of that distinction, "top down" means a layer on top of the existing Web; see further, A. Iskold, 2007. "Top Down: A New Approach to the Semantic Web," in *ReadWrite Web*, Sept. 20, 2007. The problem with this terminology is that it offers a completely different sense of "top down" to traditional uses. In Iskold's argument, his "top down" is a layering on top of the existing Web. On the other hand, "top down" is more often understood in the sense of a "comprehensive, engineered" view, consistent with [1].

[4] See M. K. Bergman, 2009. The Open World Assumption: Elephant in the Room, December 21, 2009. The open world assumption (OWA) generally asserts that the lack of a given assertion or fact being available does not imply whether that possible assertion is true or false: it simply is not known. In other words, lack of knowledge does not imply falsity. Another way to say it is that everything is permitted until it is prohibited. OWA lends itself to incremental and incomplete approaches to various modeling problems.

The closed world assumption (CWA) is a key underpinning to most standard relational data systems and enterprise schema and logics. CWA is the logic assumption that what is not currently known to be true, is false. For semantics-related projects there is a corollary problem to the use of CWA which is the need for upfront agreement on what all predicates "mean", which is difficult if not impossible in reality when different perspectives are the explicit purpose for the integration.

[5] See M.K. Bergman, 2010. "Two Contrasting Styles for the Semantic Enterprise," *AI3:::Adaptive Information* blog post, February 15, 2010. See http://www.mkbergman.com/866/two-contrasting-styles-for-the-semantic-enterprise/.

[6] I first used the term in passing in M.K. Bergman, 2007. "An Intrepid Guide to Ontologies," *AI3:::Adaptive Information* blog post, May 16, 2007. See http://www.mkbergman.com/374/an-intrepid-guide-to-ontologies/, then more fully elaborated the idea in "Where are the Road Signs for the Structured Web," *AI3:::Adaptive Information* blog post, May 29, 2007. See http://www.mkbergman.com/375/where-are-the-road-signs-for-the-structured-web/.

[7] See Catherine C. Marshall and Frank M. Shipman, 2003. "Which Semantic Web?," in *Proceedings of ACM*

Hypertext 2003, pp. 57-66, August 26-30, 2003, Nottingham, United Kingdom; http://www.csdl.tamu.edu/~marshall/ht03-sw-4.pdf, for a very different (but still accurate and useful) way to characterize the "visions" for the semantic Web. In this early paper, the authors posit three competing visions: 1) the development of standards, akin to libraries, to bring order to digital documents; this is the vision they ascribe to the W3C and has been largely adopted via use of URIs as identifiers, and languages such as RDF and OWL; 2) a vision of a globally distributed knowledge base (which they characterize as being Tim Berners-Lee's original vision, with examples being Cyc or Apple's (now disbanded) Knowledge Navigator; and 3) a vision of an infrastructure for the coordinated sharing of data and knowledge..

[8] See Claude E. Shannon's classic paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948.

[9] This reference is from the Wikipedia entry on the English language: Kister, Ken. "Dictionaries defined." Library Journal, 6/15/92, Vol. 117 Issue 11, p 43.

[10] See http://www-943.ibm.com/innovation/us/watson/related-content/toronto.html, or simply do a Web search on "watson toronto jeopardy" (no quotes).

[11] Readability is important because programmers spend the majority of their time reading, trying to understand and modifying existing source code, rather than writing new source code. Unreadable code often leads to bugs, inefficiencies, and duplicated code. It has been known for at least three decades that a few simple readability transformations can make code shorter and drastically reduce the time to understand it. See James L. Elshoff and Michael Marcotty, 1982. "Improving Computer Program Readability to Aid Modification," *Communications of the ACM*, v.25 n.8, p. 512-521, Aug 1982; see http://doi.acm.org/10.1145/358589.358596. From the Wikipedia entry on Readability.

[12] According the the Wikipedia entry on Language, there are an estimated 3000 to 6000 active human languages presently in existence.

[13] The *forest path* analogy comes from Atanas Kiryakov of Ontotext. The remaining analogies come from M.K. Bergman in his *AI3:::Adaptive Innovation* blog: "There's Not Yet Enough Backbone," May 1, 2007 (*backbone*); "The Role of UMBEL: Stuck in the Middle with you ...," May 11, 2008 (*infocline*, *scaffolding* and *docking port*); "Structure Paves the Way to the Semantic Web," May 3, 2007 (*roadmap*).

[14] Semi-automatic methods attempt to apply as much automated screening and algorithmic- or rules-based scoring as possible, and then allow the final choices to be arbitrated by humans. Fully automated systems, particularly involving natural language processing, are not yet acceptable because of (small, but) unacceptably high error rates in precision. The best semi-automated approaches handle all tasks that are rote or error-free, and then limit the final choices to those areas where unacceptable errors are still prevalent. As time goes on, more of these areas can be automated as algorithms, heuristics and methodologies improve. Eventually, of course, this may lead to fully automated approaches.
