The Shaky Semantics of the Semantic Web

by Mike Bergman - Wednesday, March 12, 2008

http://www.mkbergman.com/426/the-shaky-semantics-of-the-semantic-web/

Let's Stop Presenting a Muddled Mess to the Broader Public

Despite all of the breakthroughs of the past year, the semantic Web community at times looks determined to snatch defeat from the jaws of victory. It is ironic that the very thing the semantic Web is about -- *meaning* -- is also the very grounds of its current challenge. Sometimes we look like the gang that could not talk straight.

This semantic challenge arises from a variety of causes. Some are based on different contexts or perspectives (or sometimes lack thereof). Some is due to the sheer, tough slogging of defining and explicating a new field, its concepts, and standards and technology. And some, unfortunately, may at times arise from old-school thinking that to define or brand something is to "own" it. (Or, worse still, to overhype it and then not deliver.)

We are now in the second wave of the semantic Web. The first wave were those dozens of individuals, mostly from academia and the W3C, who have been laboring diligently on standards, tools and *language* for more than a decade. Most of the community's leaders come from this group and they are largely the stewards of the vision. The second wave, which arguably began when the iron association of RDF with XML was broken, has perhaps hundreds or thousands of members. Many are still researchers, but many are also tools builders and explicators and evangelists. Some, even, are entrepreneurs.

These two groups constitute the current semantic Web community, which is an active and increasingly visible one of blogs, reports, conferences, pragmatic tools and new companies and ventures. Financial interests and the business and technical press are also becoming more involved and active.

These two waves -- or however you want to bound the current community; frankly the definition is unimportant -- need to recognize that our communications must achieve better clarity as we make outreach and spread into the broader public. Muddled concepts, akin to the unfortunate earlier RDF-XML association, if not cleared up, can also hinder adoption.

We have a responsibility to think hard about what should be our common language and the clarity of our concepts moving forward. Let's not repeat language mistakes of the past.

We should not rush to embrace "market speak". Nor should we fear questioning current terminology or constructs where it is obviously slowing adoption or requires explanatory *legerdemain*. The semantic Web is about making meaningful connections and relationships; we should follow that same guidance for our language and conceptual metaphors.

Common language is like a pearl, with each new layer accreting upon the one before it. Current terms, definitions, acronyms, standards, and practices form the basic grain. But we are many players, and do not speak with one voice. Yet, even were we to do so, whatever we think best may not be adopted by the broader public. What is adopted as common language has a dynamic all its own. Practice -- not advocacy -- defines usage.

However, we do know that the concepts underlying the semantic Web are both foreign and difficult for the broader public. We can take solace that with HTML and other standards and protocols of the Web that such difficulties are not ultimately barriers to adoption. If it has value (and all of us know the semantic Web does), it will be adopted. But, on the other hand, insofar as our language is unnecessarily technical, or perhaps muddled conceptually or difficult to convey, we will unfortunately see a slower rate of adoption.

What we have is good -- indeed very good -- but it could be much better. And it is more often than not language than ideas that get in the way.

The 'Big Picture' is Not a Snapshot

The casual observer of the semantic Web can not help but see the occasional storms that roil our blogs and mailing lists. The storm activity has been especially high recently, and it has been a doozy. It has been as fundamental as defining what we are about and our space to heated flashpoints around what had seemed to be settled concepts (we'll address the latter in a bit).

The first challenge begins with how we even name our collective endeavor. For a decade, this name has been the 'Semantic Web'. But, either due to real or perceived past disappointments or the imperatives of a new era, this label has proven wanting to many. Benjamin Nowack recently compiled some of the leading alternatives being promulgated to define the Semantic Web space:

- Semantic Web (timbl)
- Web of Data (timbl)
- lowercase semantic [wW]eb (tantek)
- Semantic Web 2.0 (by stefandecker, IIRC)
- Web 3.0 (by nova and others)
- Semantic Graph (by nova and others)
- Hyperdata (by danja)
- Linked Data (by timbl, as implemented by Chris Bizer and Richard Cyganiak for Linking Open Data Community)
- Linked Data Web (by kidehen)
- Structured Web (by mkbergman)

- Semantic Data Web (by kidehen)
- SemWeb (by the developer community)
- GGG - The Giant Global Graph (by timbl)
- Web 3G (by iand).

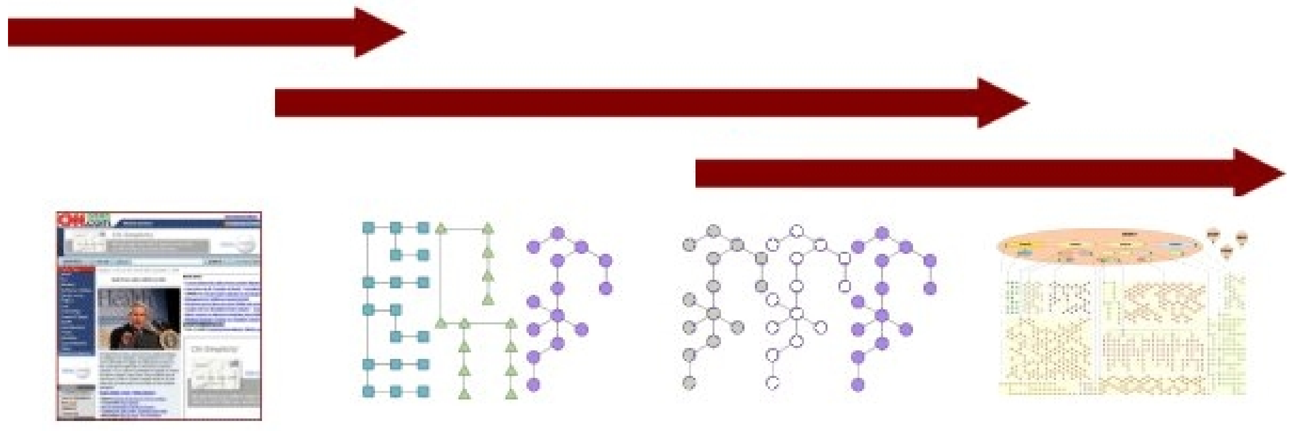
This is a useful list; but it says more than the breadth of its compilation. There is a logical flaw in trying to define the semantic Web as either a "thing" or as a static achievement. These important distinctions get swamped in the false usefulness of a list.

For example, my name is associated with one of those names on that list, the *structured Web*, but I never intended it as an alternative or marketing replacement to the semantic Web. I also have been a vocal advocate for the Linked Data concept.

We, as its current ambassadors and advocates, need to stress two aspects of the semantic Web to the broader public. First, the semantic Web is inherently about *meaning*. If we ourselves are not doing all we can to explicate the *meaning* of the semantic Web in our language and terminology, then we are doing the vision and our responsibilities a disservice.

Second, the semantic Web is also an ideal, a vision perhaps, that will not appear as a bright line in the dark or a flash of light from the blue. It will take time and much effort to approximate that vision, and its ultimate realization is as unlikely as the timeless human desire for universal communication as captured by the apocryphal Tower of Babel.

Today's circumstance is not a competition for a static label, but a constant evolution from unstructured to structured data and then increasing meaning about that data. Our true circumstance today -- that is, the current "state" of this evolution to the semantic Web -- is Linked Data, as placed in this diagram:[1]



Document Web	Structured Web	Semantic Web Linked Data	
Document-centric	Data-centric	Data-centric	Data-centric
Document resources	Structured data	Linked data	Linked data
Unstructured data and semi-structured data	Semi-structured data and structured data	Semi-structured data and structured data	Semi-structured data and structured data
HTML	XML, JSON, RDF,	RDF, RDF-S	RDF, RDF-S, OWL

etc

- URL-centric
  - *circa* 1993
- URI-centric
  - *circa* 2003
- URI-centric
  - *circa* 2007
- URI-centric
  - *circa* ???

Many of the so-called alternative terms for the semantic Web are really attempts to capture this dynamic of evolution and development. On the other hand, watch out for those terms that try to "brand" or label the semantic Web as a static event; they are the ones that lack *meaning*.

It is a cliché that conflict sells newspapers. But, we also know that newspaper sales are dropping daily. Old thinking that tries to "brand" the semantic Web space is ultimately due to fail because it is fundamentally at odds with the purpose of the semantic Web itself -- linking relevant information together with meaning. Meaningless labels are counter to this aim.

Aside from the hijackers, the community itself, sure, should want better language and communications. That, after all, is the purpose of this post. But, there is nothing to be ashamed of with the banner of the 'semantic Web', the original and still most integral view of Tim Berners-Lee. Let's just be clear this is not a bright line achievement, it is an ongoing process, and that there may be many labels that effectively capture the evolution of the vision as it may then be current at any point in time.

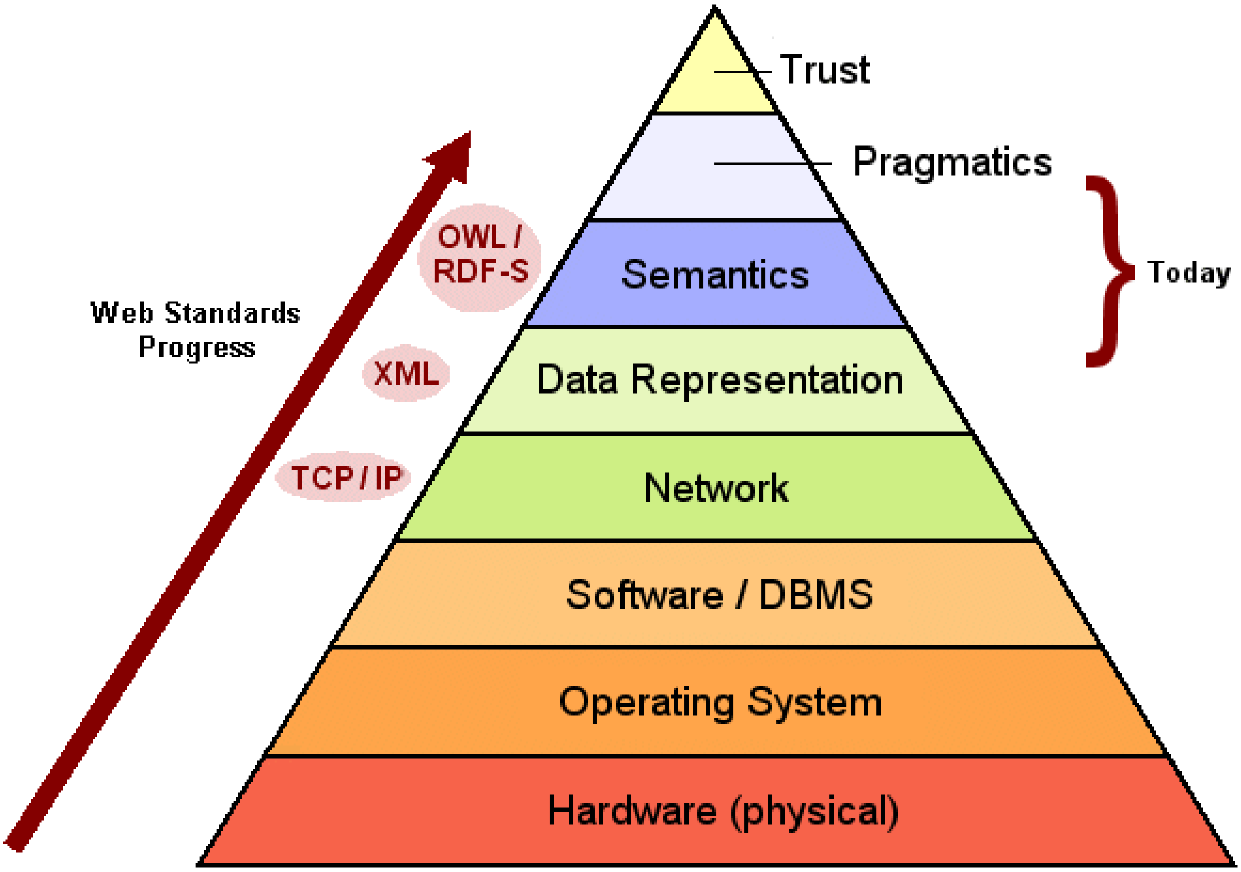
Today, that state of the art is Linked Data.

Data Federation in Perspective

It is easy to forget the recent past. To appreciate the current state of the semantic Web and its prospects, it is essential to understand the progress in data federation over the past decade or two. Incredible progress has been made to overcome what had been perceived as close-to intractable data interoperability issues within my working lifetime.[2]

"Data federation" -- the recognition that value could be unlocked by connecting information from multiple, separate data stores -- became a research emphasis within the biology and computer science communities in the 1980s. It also gained visibility as "data warehousing" within enterprises by the early-90s. However, within that period, diversity and immaturity in hardware, operating systems, databases, software and networking hampered the sharing of data.

From the perspective of 25 years ago, when we were at the bottom, the "data federation" challenge was an imposing pyramid:



Rapid Progress in Climbing the Data Federation Pyramid

When the PC came on the scene, there were mainframes from weird 36-bit Data General systems to DEC PDP minicomputers to the PCs themselves. Even on PCs, there were multiple operating systems, and many then claimed that CP/M was likely to be ascendant, let alone the upstart MS-DOS or the gorilla threat of OS/2. Hardware differences were all over the map, operating systems were a laundry list two pages long, and nothing worked with anything else. Computing in that era was an Island State.

Client-server and all of the "N-tier" speak soon followed, and it was sort of an era of progress but still costly and proprietary answers to get things to talk to one another. Yet there was beginning to emerge a rationality, at least at the enterprise level, for how to link resources together from the mainframe to the desktop. Computing in that era was the Nation State.

But still, it was incredibly difficult to talk with other nations. And that is where the Internet, specifically the Web protocol and the Mosaic (later Netscape) browser came in. Within five years (actually less) from 1994 onward the Internet took off like a rocket, doubling in size every 3-6 months.

In the early years of trying to find standards for representing semi-structured data (though not yet called that), the major emphasis was on data transfer protocols. In the financial realm, one standard dating from the late 1970s was electronic data interchange (EDI). In science, there were literally tens of exchange forms proposed with varying degrees of acceptance, notably abstract syntax notation (ASN.1), TeX (a typesetting system created by Donald Knuth and its variants such as LaTeX), hierarchical data format

(HDF), CDF (common data format), and the like, as well as commercial formats such as Postscript and others.

Of course, midway into these data representation efforts was the shift to the Internet Age, blowing away many previous notions and limits. The Internet and its TCP/IP protocols and HTML and XML standards for "semi-structured" data and data transfer and representations, in particular, overcame physical and syntactical and data exchange heterogeneities, helping us climb the data federation pyramid still higher.

Yet, as the pyramid shows, despite massive progress in scaling it, challenges remain now centered at the layer of semantics. Related to that are the issues of what data you can trust and with what authority.

A Rose by Any Other Name

Now that the focus has evolved to the level of *meaning*, the standards work of the semantic Web of the past ten years faces the challenge of being communicated and inculcated into the broader public. Frankly, this is proving to be rough sledding, and is one of the reasons, perhaps, why others are seeking new marketing terms and to diminish the semantic Web label.

To be sure, as with any new field over history from agriculture to rocket science, there is technical terminology associated with the semantic Web. Things such as RDF, OWL, ontologies, SPARQL, GRDDL, and many, many other new terms have been discussed and presented in beginning guides elsewhere (including by me). Such specificity and terminology is expected and natural in new technical areas. In itself, new terminology is not automatically grounds for hindering adoption.

But, in a growing crescendo on semantic Web mailing lists, and to which I agree in some measure, some of the conceptual and terminological underpinnings of the semantic Web are being shaken hard.

As background, realize that the subjects and relationships available on the semantic Web are defined as requiring a unique identifier, one which corresponds to the addressing scheme of the Web now familiar to us as URLs. However, since URLs were designed for single documents or Web pages, they now need to be abstracted a level further to a URI to reflect the potential data objects formerly masked within a document-centric URL view.

In the initial Web, a document was a document, and a URL could clearly be related to its representation. But, now that the resource objects are shifting from documents to data, a document could potentially contain multiple data items, or could be a reference to the actual data elsewhere. This potential indirection has proven to be a real bugaboo to semantic Web researchers.

The authors of the semantic Web are mostly computer scientists and not linguists or semanticists. They have looked at the idea of resources quite literally, and have tended to view whether the target of a reference is to its actual document object (an "information resource", linked to via a URL) versus an indirection source (a "non-information resource", referred to by its URI). This tenuous distinction between information and non-information appears arbitrary and, in any case, is difficult for most to understand. (It also fundamentally confuses the notions of resource and representation.)

Unfortunately, this same literal perspective has tended to perpetuate itself at other levels of confusion.

Attempting to access a "non-information" resource forces the need to resolve to an actual retrieval address, a process called "dereferencing". Lacking true semantic sophistication, some resources felt to represent the same object (but lacking any metrics for actually determining this equivalence) have also been related to one another through (in some instances) an indefensible "same as" relationship. URIs may also be absent for subject or object nodes in an RDF triple causing an empty reference in those instances and so-called blank nodes, or "bnodes."

The power of the RDF triple easily trumps these shortcomings, but the shortcomings remain nonetheless. These shortcomings manifest themselves at times in wrong conceptual underpinnings or relationships or at other times in making communication of the system very difficult to newcomers.

Most of these problems can be readily fixed. RDF itself is a very solid foundation and has no real shortcomings at this point. The remaining problems are mostly terminological and where logic changes might be warranted, those are minor.

There has, however, been keen reluctance to recognize these logical and terminology shortcomings. Those failures are somewhat an artifact of the underlying semantics. The Web and the Internet, of course, is a representational system, and not one of true resources or objects. That hang up alone has cost the community major bonus bucks in getting its story and terminology straight.

Alas, while the system works splendidly, it sorely needs true semantic attention from linguists, philosophers and ontologists that better understand representation *v.* resources. This is a missing step, that, if taken, could truly be the missing secret sauce.

Lastly, there are parties trying to coin new terms and terminology in order to "own" the evolving semantic Web. This was a strategy that worked in decades past for enterprise systems where vendors tried to define new "spaces". But, in the context of the semantic Web where the objective is interoperability and not proprietary advantage, such approaches are doomed to failure, as well as unfortunately acting as brakes on broader adoption.

The Web is a Dirty Place

Semantic Web terminology nuances and subtleties that dance on the head of pins moreover belie another fundamental truth: The Web is a dirty place. If users or automatic software can screw it up, it will be screwed up. If darker forces can find bypasses and trapdoors, those will be opened as well.

This should be no shocking truth to anyone who has worked on the Web for any period longer than a few years. In fact, in one of my former lives, we made our money by exploiting that very diversity and imperfect chaos by finding ways to talk to hundreds of thousands of dynamic search forms and databases. Talk about screwed up! Every search engine vendor and any Web site developer who has to struggle with getting CSS to work cross-browser knows intimately at least in part of what I speak.

It is thus perplexing how many in the semantic Web community -- who should truly know better -- have continued to promote and advocate systems that are fragile, easily broken, and not robust. While RDF is certainly the best *lingua franca* for data interoperability, it is incredibly difficult to set up Web sites to publish in that format with proper dereferencing and exposure of URIs and endpoints. The simpler protocols of the standard Web public and standard Web developers reflect adoption realities that should not be dismissed, but actively embraced and co-opted. Denials and either-or arguments set up artificial boundaries and limit community growth and acceptance.

Ten-page how-to manuals for how to publish semantic Web data and complicated stories about 303 see other redirects and Apache technical configurations are a loser. This is not the way to gain broad-scale adoption. And it certainly is not a great way to make new friends.

I'm no technogeek and I have in good faith struggled myself to adhere to these techniques. I frankly still don't get it and wonder how many others ever will, let alone have any interest or desire to do so. Anything that is harder than writing a Web page is too hard. Tasks need to be approachable in bit-size chunks and in true REST style.

There have been some very important voices of reason speaking to these issues on the specialty mailing lists in the semantic Web community, but I perceive denial and an unwillingness to engage a meaningful dialog and quick resolution to these matters. Community, I can not say this more clearly: Wake up! It is time for a mature acceptance of reality to set in and to get on with the real task of solving problems within the context of the real Web and its denizens.

Meaningful Semantics is Tougher Than it Looks

The flip side to this unnecessary complexity is an almost childlike simplicity to what the true "semantics" of the semantic Web really *means*. Standard current arguments tend to be framed about whether the same person or named entity in one representation is the same as or somehow different than other representations.

Granted, this is but a continuation of the resource *v.* representation issue earlier mentioned. And, while the community gets diverted in such fruitless and unnecessary debates, real semantic complexity gets overlooked. Yet these are complex and vexing boils just ready to erupt through the surface.

This strange fixation on topics that should be easily resolved and lack of attention to real semantic topics is perhaps a function of the community's heritage. Most semantic Web researchers have a background in computer science. Linguists, semioticians, philosophers, ontologists, and disciplines with real semantic training and background certainly are active in the community, but are greatly outnumbered.

A short story will illustrate this point. About two years ago I began collecting a listing of as many semantic Web tools as I could find, with the result that Sweet Tools has become one of the community's go-to listings. But, shortly after I began publishing the list, while getting nice compliments from the W3C's semantic Web tools site, it was also noted in passing that my site "also lists tools that are *related* to Semantic Web though not necessarily using the core technology (*e.g.*, natural language ontology tools). . . ."[3] If resolving and understanding the *meaning* of language is not at the core of the semantic Web, it is hard to know what is.

Frankly, it is this *semantics* aspect that will push real progress on the semantic Web back into the future. While Linked Data is showing us the technical aspects for how we can bring disparate data together, and ontologies will help us place that data into the right ballpark -- all of which is and will bring real benefits and value -- much work still remains.

Semantic mediation -- that is, resolving semantic heterogeneities -- must address more than 40 discrete categories of potential mismatches from units of measure, terminology, language, and many others. These sources may derive from structure, domain, data or language. [4] Possible drivers in semantic mismatches can occur from world view, perspective, syntax, structure and versioning and timing:

- One schema may express a similar world view with different syntax, grammar or structure
- One schema may be a new version of the other
- Two or more schemas may be evolutions of the same original schema
- There may be many sources modeling the same aspects of the underlying domain ("horizontal resolution" such as for competing trade associations or standards bodies), or
- There may be many sources that cover different domains but overlap at the seams ("vertical resolution" such as between pharmaceuticals and basic medicine).

These differences in purpose and provenance are the sources of these mismatches. Pluempitiwiriyawej and Hammer classify heterogeneities into three broad classes:[5]

- *Structural* conflicts arise when the schema of the sources representing related or overlapping data exhibit discrepancies. Structural conflicts can be detected when comparing the underlying ontologies. The class of structural conflicts includes generalization conflicts, aggregation conflicts, internal path discrepancy, missing items, element ordering, constraint and type mismatch, and naming conflicts between the element types and attribute names.
- *Domain* conflicts arise when the semantics of the data sources that will be integrated exhibit discrepancies. Domain conflicts can be detected by looking at the information contained in the ontologies and using knowledge about the underlying data domains. The class of domain conflicts includes schematic discrepancy, scale or unit, precision, and data representation conflicts.
- *Data* conflicts refer to discrepancies among similar or related data values across multiple sources. Data conflicts can only be detected by comparing the underlying actual data. The class of data conflicts includes ID-value, missing data, incorrect spelling, and naming conflicts between the element contents and the attribute values.

Moreover, mismatches or conflicts can occur between set elements (a "population" mismatch) or attributes (a "description" mismatch).

The table below thus builds on their schema by adding the fourth major explicit category of language, leading to about 40 distinct potential sources of semantic heterogeneities:

Class	Category	Subcategory
STRUCTURAL	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	Generalization / Specialization	
	Aggregation	Intra-aggregation
		Inter-aggregation

	Internal Path Discrepancy	
	Missing Item	Content Discrepancy
		Attribute List Discrepancy
		Missing Attribute
		Missing Content
	Element Ordering	
	Constraint Mismatch	
	Type Mismatch	
DOMAIN	SchematicDiscrepancy	Element-value to Element-label Mapping
		Attribute-value to Element-label Mapping
		Element-value to Attribute-label Mapping
		Attribute-value to Attribute-label Mapping
	Scale or Units	
	Precision	
	DataRepresentation	Primitive Data Type
		Data Format
DATA	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	ID Mismatch or Missing ID	
	Missing Data	
	Incorrect Spelling	
LANGUAGE	Encoding	Ingest Encoding Mismatch
		Ingest Encoding Lacking
		Query Encoding Mismatch
		Query Encoding Lacking
	Languages	Script Mismatches
		Parsing / Morphological Analysis Errors (many)
		Syntactical Errors (many)
		Semantic Errors (many)

Most of these line items are self-explanatory, but a few may not be. See further [4].

Mature and accepted ontologies -- largely lacking in most topics and domains -- will be key sources to help overcome some of these heterogeneities, but are still quite far off. Yet ontologies alone will never be complete enough to resolve all mismatches.

Some Respectful Recommendations

This whirlwind tour of the state of the semantic Web shows tremendous progress and tremendous efforts still to come. The vision of the semantic Web is certainly one of a process -- an ongoing journey -- and one that defies facile labels. Further, there is real *meaning* behind the semantic Web and its striving to find meaning in heretofore disconnected and disparate data.

This is a journey of substance and value. It is one that is exciting and challenging and rewarding. As someone once joked to me, "Welcome to the multi-generational, full-employment act!"

The value to be gained from the semantic Web enterprise is such that perhaps we can not avoid the hucksters and hypsters and spinmeisters. At times there seems to be a chilly wind of big bucks and excessive VC froth blowing through the community. I guess the best advice is to stay attuned to where real meaning and substance occurs, and hold your wallet in the other parts of town.

We as a community could be doing better, however, in the nature and language of the substance that we do offer to the broader public. With that aim, let me offer some thoughts on immediate steps we can take to promote that objective:

- The community's ultimate banner needs to be Berner-Lee's Semantic Web vision. Whether we use upper or lower case does not matter. *The Economist* magazine, for example, now lowercases the Web and Internet; these are journalistic and style differences ultimately. Talking about upper and lower case Semantic Web starts to look silly in that light though personally I use semantic Web (because I still capitalize Web and Internet! and don't like to convey total upper case "states"). But, hey, that is my own preference and ultimately who cares?
- As a community, we should try diligently to convey the semantic Web as a process and not a fixed event or stage of the Web. Insofar as terms and labels can help the public understand our current state and capabilities, use all of the terms and labels within our language to communicate and convey meaning under this umbrella. In any event, let us abandon silly version numbers that mean nothing and embrace the idea of process and stages, not fixed realities
- Let us pledge to not try to introduce language and terminology for branding, hype and "ownership" reasons. Not only does it fail, but it undercuts the entire community's prospects and timing of when it will be successful. Just as our community has moved to open standards, open source and open data, realize that hype and branding crap is closed crap, and counter to our collective interest, and
- We need to abandon or revise the earlier layer cake by removing its prominent role for XML.

Finally, if they would accept it (though it is presumptuous of me since I don't know them and have not asked them), have the W3C ask Pat Hayes and Roy Fielding to work out the resource and representational terminology issues. Perhaps we need a bit of benevolent dictatorship at this point rather than more muddling by committee.

Besides finding I personally agree with and like most of what these two write, Pat is a great choice because he is the author of the best description of the semantic underpinnings of RDF [6] and Roy is a great choice because he clearly understands the representational nature of the Web architecture as exemplified in his REST thesis [7]. The time is now and sorely needed to get the issues of representation, resources and reference cleaned up once and for all. The W3C TAG, though dedicated and obviously well-intentioned, has arguably not helped matters in these regards. I would be gladly willing to give Pat and Roy my proxy (assuming I had one :)) on issues of terminology and definitions.

[1] Last July I wrote a piece entitled, More Structure, More Terminology and (hopefully) More Clarity. It, and related posts on the structured Web, had as its thesis that the Web was naturally evolving from a document-centric basis to a "Web of Data". We already have much structured data available and the

means through RDFizers and other techniques to convert that structure to Linked Data. Linked Data thus represented a very doable and pragmatic way station on the road to the semantic Web. It is a journey we can take today; indeed, many already are as growth figures attest

.[2] V.M. Markowitz and O. Ritter, "Characterizing Heterogeneous Molecular Biology Database Systems," in *Journal of Computational Biology* 2(4): 547-546, 1995.

[3] That statement has changed in nuance a number of times over the months, and was finally removed from the site in about January 2008.

[4] Much of this section was drawn from a posting by me on Sources and Classification of Semantic Heterogeneities in June 2006.

[5] Charnyote Pluempitiwiriyawej and Joachim Hammer, "A Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources," *Technical Report TR00-004*, University of Florida, Gainesville, FL, 36 pp., September 2000. See ftp.dbcenter.cise.ufl.edu/Pub/publications/tr00-004.pdf.

[6] Patrick Hayes, ed., 2004. *RDF Semantics*, W3C Recommendation 10 February 2004. See http://www.w3.org/TR/rdf-mt/.

[7] Roy T. Fielding, 2000. *Architectural Styles and the Design of Network-based Software Architectures*, doctoral thesis, Department of Information and Computer Science, University of California at Irvine, 179 pp. See http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation_2up.pdf

PDF generated by *AI3:::Adaptive Information* blog