The Bipolar Disorder of Linked Data

by Mike Bergman - Wednesday, April 28, 2010

http://www.mkbergman.com/880/the-bipolar-disorder-of-linked-data/

An Acceptance of Its Natural Role is the Prozac Substitute

There has been a bit of a manic-depressive character on the Web waves of late with respect to linked data. On the one hand, we have seen huzzahs and celebrations from the likes of ReadWriteWeb and Semantic Web.com and, just concluded, the Linked Data on the Web (LDOW) workshop at WWW2010. This treatment has tended to tout the coming of the linked data era and to seek ideas about possible, cool linked data apps [1]. This rise in visibility has been accomplished by much manic and excited discussion on various mailing lists.

On the other hand, we have seen much wringing of hands and gnashing of teeth for why linked data is not being used more and why the broader issue of the semantic Web is not seeing more uptake. This depressive "call to arms" has sometimes felt like ravings with blame being given to the poor state of apps and user interfaces to badly linked data to the difficulty of publishing same. Actually using linked data for anything productive (other than single sources like DBpedia) still appears to be an issue.

Meanwhile, among others, Kingsley Idehen, ubiquitous voice on the Twitter #linkeddata channel, has been promoting the separation of identity of linked data from the notion of the semantic Web. He is also trying to change the narrative away from the association of linked data with RDF, instead advocating "Data 3.0" and the entity-attribute-value (EAV) model understanding of structured data.

As someone less engaged in these topics since my own statements about linked data over the past couple of years [2], I have my own distanced-yet-still-biased view of what all of this crisis of confidence is about. I think I have a diagnosis for what may be causing this bipolar disorder of linked data [3].

The Semantic Web Boogie Man

A fairly universal response from enterprise prospects when raising the topic of the semantic Web is, "That was a big deal of about a decade ago, wasn't it? It didn't seem to go anywhere." And, actually, I think both proponents and keen observers agree with this general sentiment. We have seen the original advocate, Tim Berners-Lee, float the Giant Global Graph balloon, and now Linked Data. Others have touted Web 3.0 or Web of Data or, frankly, dozens of alternatives. Linked data, which began as a set of techniques for publishing RDF, has emerged as a potential marketing hook and saviour for the tainted original semantic Web term.

And therein, I think, lies the rub and the answer to the bipolar disorder.

If one looks at the original principles for putting linked data on the Web or subsequent interpretations, it is clear that linked data (lower case) is merely a set of techniques. Useful techniques, for sure; but really a simple approach to exposing data using the Web with URLs as the naming convention for objects and their relationships. These techniques provide (1) methods to access data on the Web and (2) specifying the relationships to link the data (resources). The first part is mechanistic and not really of further concern here. And, while any predicate can be used to specify a data (resource) relationship, that relationship should also be discoverable with a URL (dereferencable) to qualify as linked data. Then, to actually be semantically useful, that relationship (predicate) should also have a precise definition and be part of a coherent schema. (Note, this last sentence is actually not part of the "standard" principles for linked data, which itself is a problem.)

When used right, these techniques can be powerful and useful. But, poor choices or execution in how relationships are specified often leads to saying little or nothing about semantics. Most linked data uses a woefully small vocabulary of data relationships, with even a smaller set ever used for setting linkages ***across*** existing linked data sets [4]. Linked data techniques are a part of the foundation to overall best practices, but not the total foundation. As I have argued for some time, linked data alone does not speak to issues of context nor coherence.

To speak semantically, linked data is not a synonym for the semantic Web nor is it the `sameAs` the semantic Web. But, many proponents have tried to characterize it as such. The general tenor is to blow the horns hard anytime some large data set is "exposed" as linked data. (No matter whether the data is incoherent, lacks a schema, or is even poorly described and defined.) Heralding such events, followed by no apparent usefulness to the data, causes confusion to reign supreme and disappointment to naturally occur.

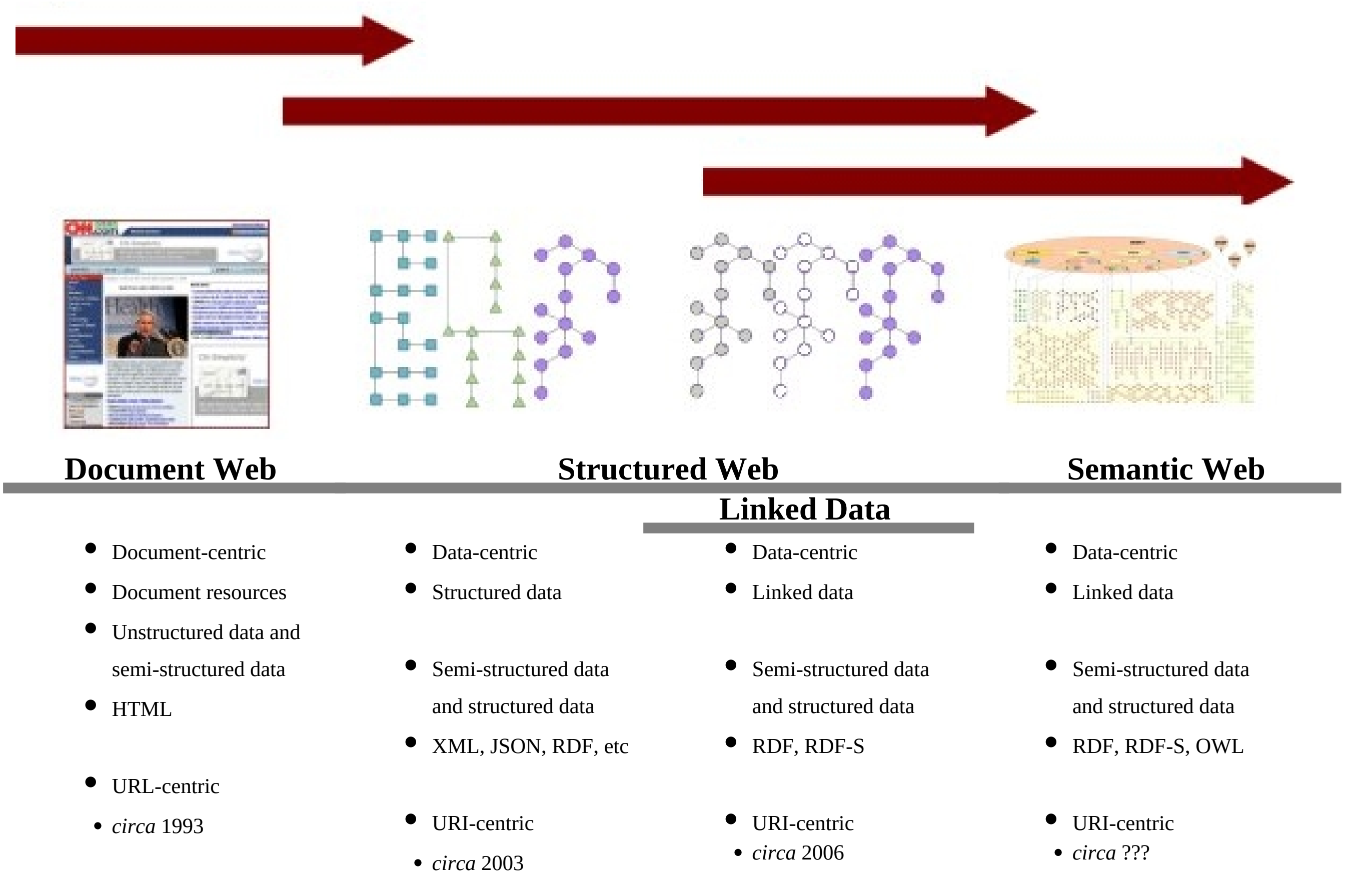
The semantic Web (or semantic enterprise or semantic government or similar expressions) is a vision and an ideal. It is also a fairly complete one that potentially embraces machines and agents working in the background to serve us and make us more productive. There is an entire stack of languages and techniques and methods that enable schema to be described and non-conforming data to be interoperated. Now, of course this ideal is still a work in progress. Does that make it a failure?

Well, maybe so, if one sees the semantic Web as marketing or branding. But, who said we had to present it or understand it as such?

The issue is not one of marketing and branding, but the lack of benefits. Now, maybe I have it all wrong, but it seems to me that the argument needs to start with what "linked data" and the "semantic Web" can do for me. What I actually call it is secondary. Rejecting the branding of the semantic Web for linked data or Web 3.0 or any other somesuch is still dressing the emperor in new clothes.

A Nicely Progressing Continuum, Thank You!

For a couple of years now I have tried in various posts to present linked data in a broader framework of structured and semantic Web data. I first tried to capture this continuum in a diagram from July 2007:



Now, three years later, I think the transitional phase of linked data is reaching an end. OK, we have figured out one useful way to publish large datasets staged for possible interoperability. Sure, we have billions of triples and assertions floating out there. But what are we to do with them? And, is any of it any good?

The Reality of a Heterogeneous World

I think Kingsley is right in one sense to point to EAV and structured data. We, too, have not met a structured data format we did not like. There are hundreds of attribute-value pair models of even more generic nature that also belong to the conversation.

One of my most popular posts on this blog has been, ‘Structs’: Naïve Data Formats and the ABox, from January 2009. Today, we have a multitude of popular structured data formats from XML to JSON and even spreadsheets (CSV). Each form has its advocates, place and reasons for existence and popularity (or not). This inherent diversity is a fact and fixture of any discussion of data. It is a major reason why we developed the irON (*instance record* and *object notation*) non-RDF vocabulary to provide a bridge from such forms to RDF, which is accessible on the Web via URIs. irON clearly shows that entities can be usefully described and consumed in either RDF or non-RDF serialized forms.

Though RDF and linked data is a great form for expressing this structured information, other forms can convey the same meaning as well. Of the billions of linked data triples exposed to date, surely more than 99% are of this instance-level, "ABox" type of data [5]. And, more telling, of all of the structured data that is publicly obtainable on the Web, my wild guess is that less than 0.0000000001% of that is even linked RDF data [6].

Neither linked data nor RDF alone will -- today or in the near future -- play a pivotal or essential role for instance data. The real contribution from RDF and the semantic Web will come from connecting things together, from interoperation and federation and conjoining. This is the provenance of the TBox and is a role barely touched by linked data. Publishing data as linked data helps tremendously in simplifying ingest and guiding the eventual connections, but the making of those connections, testing for their quality and reliability, are steps beyond the linked data ken or purpose.

Promoting Linked Data to its Level of Incompetence

It seems, then, that we see two different forces and perspectives at work, each contributing in its own way to today's bipolar nature of linked data.

On the manic side, we see the celebration for the release of each large, linked data set. This perspective seems to care most about volumes and numbers, with less interest in how and whether the data is of quality or useful. This perspective seems to believe "post the data, and the public will come." This same perspective is also quite parochial with respect to the unsuitability of non-linked data, be it microdata, microformats or any of the older junk.

On the depressed side, linked data has been seen as a more palatable packaging for the disappointments and perceived failures or slow adoption of the earlier semantic Web phrasing. When this perspective sees the lack of structure, defensible connections and other quality problems with linked data as it presently exists, despair and frustration ensue.

But both of these perspectives very much miss the mark. Linked data will never become the universal technique for publishing structured data, and should not be expected to be such. Numbers are never a substitute for quality. And linked data lacks the standards, scope and investment made in the semantic Web to date. Be patient; don't despair; structured data and the growth of semantics and useful metadata is proceeding just fine.

Unrealistic expectations or wrong roles and metrics simply confuse the public. We are fortunate that most potential buyers do not frequent the community's various mailing lists. Reduced expectations and an understanding of linked data's natural role is perhaps the best way to bring back balance.

Linked Data's Natural Role

We have consciously moved our communications focus from speaking internally to the community to reaching out to the broader enterprise public. There is much of education, clarification and dialog that is now needed with the buying public. The time has moved past software demos and toys to workable, pragmatic platforms, and the methodologies and documentation necessary to support them. This particular missive speaking to the founding community is (perhaps many will Hurray!) likely to become even more rare as we continue to focus outward.

As Structured Dynamics has stated many times, we are committed to linked data, presenting our information as such, and providing better tools for producing and consuming it. We have made it one of the seven foundations to our technology stack and methodology.

But, linked data on its own is inadequate as an interoperability standard. Many practitioners don't publish it right, characterize it right, or link to it right. That does not negate its benefits, but it does make it a poor candidate to install on the semantic Web throne.

Linked data based on RDF is perhaps the first citizen amongst all structured data citizens. It is an expressive and readily consumed means for publishing and relating structured instance data and one that can be easily interoperated. It is a natural citizen of the Web.

If we can accept and communicate linked data for these strengths, for what it naturally is -- a useful set of techniques and best practices for enabling data that can be easily consumed -- we can rest easy at night and not go crazy. Otherwise, bring on the Prozac.

---

[1] Actually, in my opinion, the suggested listing of apps from these discussions is distinctly unimpressive and not compelling. As argued in the main body of the post, I think this is because linked data is really just a technique or best practice, and not a basis alone for enabling compelling apps. As initial developers of such apps as the UMBEL concept explorer or Dataviewer, Structured Dynamics understands the use of linked data and has a defensible basis to comment on applications. Our own applications intimately integrate linked data, but only as one of seven foundations.

[2] Here are some of my relevant posts over the past year discussing the role of linked data: Moving Beyond Linked Data (Sept. 20, 2009); Fresh Perspectives on the Semantic Enterprise (Sept. 28, 2009); The Law of Linked Data (Oct. 11, 2009); When Linked Data Rules Fail (Nov. 16, 2009).

[3] The current bipolar discussion reminds me of the "Six Phases of a Project," a copy of which has been a permanent fixture on my office wall:

1. Enthusiasm
2. Disillusionment
3. Panic
4. Search for the guilty
5. Punishment of the innocent
6. Honors & praise for the non-participants.

[4] See, for example: Harry Halpin, 2009. "A Query-Driven Characterization of Linked Data," paper presented at the Linked Data on the Web (LDOW) 2009 Workshop, April 20, 2009, Madrid, Spain, see http://events.linkeddata.org/ldow2009/papers/ldow2009_paper16.pdf; Prateek Jain, Pascal Hitzler, Peter Z. Yehy, Kunal Vermay and Amit P. Shet, 2010. "Linked Data is Merely More Data," in Dan Brickley, Vinay K. Chaudhri, Harry Halpin, and Deborah McGuinness, *Linked Data Meets Artificial Intelligence*, *Technical Report SS-10-07*, AAAI Press, Menlo Park, California, 2010, pp. 82-86., see http://knoesis.wright.edu/library/publications/linkedai2010_submission_13.pdf; among others.

[5] Structured Dynamics' best practices approach makes explicit splits between the “ABox” (for instance data) and “TBox” (for ontology schema) in accordance with our working definition for description logics, a fundamental underpinning for how we use RDF:

“Description logics and their semantics traditionally split *concepts* and their relationships from the different treatment of *instances* and their attributes and roles, expressed as fact assertions. The concept split is known as the TBox (for *terminological* knowledge, the basis for *T* in *TBox*) and represents the schema or taxonomy of the domain at hand. The TBox is the structural and intensional component of conceptual relationships. The second split of instances is known as the ABox (for *assertions*, the basis for *A* in *ABox*) and describes the attributes of instances (and individuals), the roles between instances, and other assertions about instances regarding their class membership with the TBox concepts.”

[6] This topic is deserving of some analysis in its own right, and my guess is really just that. For example, RSS feeds to mobile devices alone perhaps account for 2,000 petabytes today; see http://www.tgdaily.com/hardware-features/49167-8000-petabytes-of-mobile-data-traffic-expected-by-2014.