# I Have Yet to Metadata I Didn't Like

**by Mike Bergman - Monday, August 16, 2010**

http://www.mkbergman.com/902/i-have-yet-to-metadata-i-didnt-like/

## Contrasted with Some Observations on Linked Data

At the SemTech conference earlier this summer there was a kind of vuvuzela-like buzzing in the background. And, like the World Cup games on television, in play at the same time as the conference, I found the droning to be just as irritating.

That droning was a combination of the sense of righteousness in the superiority of linked data matched with a reprise of the "chicken-and-egg" argument that plagued the early years of semantic Web advocacy [1]. I think both of these premises are misplaced. So, while I have been a fan and explicator of linked data for some time, I do not worship at its altar [2]. And, for those that do, this post argues for a greater sense of ecumenism.

My main points are not against linked data. I think it a very useful technique and good (if not best) practice in many circumstances. But my main points get at whether linked data is an objective in itself. By making it such, I argue our eye misses the ball. And, in so doing, we miss making the connection with ***meaningful, interoperable information***, which should be our true objective. We need to look elsewhere than linked data for root causes.

### Observation #1: What Problem Are We Solving?

When I began this blog more than five years ago -- and when I left my career in population genetics nearly three decades before that -- I did so because of my belief in the value of information to confer adaptive advantage. My perspective then, and my perspective now, was that adaptive information through genetics and evolution was being uniquely supplanted within the human species. This change has occurred because humanity is able to record and carry forward all information gained in its experiences.

Adaptive innovations from writing to bulk printing to now electronic form uniquely position the human species to both record its past and anticipate its future. We no longer are limited to evolution and genetic information encoded in surviving offspring to determine what information is retained and moves forward. Now, ***all*** information can be retained. Further, we can combine and connect that information in ways that break to smithereens the biological limits of other species.

Yet, despite the electronic volumes and the potentials, chaos and isolated content silos have characterized humanity's first half century of experience with digital information. I have spoken before about how we have been steadily climbing the data federation pyramid, with Internet technologies and the Web being prime factors for doing so. Now, with a compelling data model in RDF and standards for how we can relate any type of information meaningfully, we also have the means for making sense of it. And connecting it. And learning and adapting from it.

And, so, there is the answer to the rhetorical question: The problem we are solving is to ***meaningfully connect information***. For, without those meaningful connections and recombinations, none of that information confers adaptive advantage.

## Observation #2: The Problem is Not A Lack of Consumable Data

One of the "chicken-and-egg" premises in the linked data community is there needs to be more linked data exposed before some threshold to trigger the network effect occurs. This attitude, I suspect, is one of the reasons why hosannas are always forthcoming each time some outfit announces they have posted another chunk of triples to the Web.

Fred Giasson and I earlier tackled that issue with When Linked Data Rules Fail regarding some information published for data.gov and the New York Times. Our observations on the lack of standards for linked data quality proved to be quite controversial. Rehashing that piece is not my objective here.

What ***is*** my objective is to hammer home that we do not need linked data in order to have data available to consume. Far from it. Though linked data volumes have been growing, I actually suspect that its growth has been slower than data availability *in toto*. On the Web alone we have searchable deep Web databases, JSON, XML, microformats, RSS feeds, Google snippets, yada, yada, all in a veritable deluge of formats, contents and contexts. We are having a hard time inventing the next 1000-fold description beyond zettabyte and yottabyte to even describe this deluge [3].

There is absolutely no voice or observer anywhere that is saying, "We need linked data in order to have data to consume." Quite the opposite. The reality is we are drowning in the stuff.

Furthermore, when one dissects what most of all of this data is about, it is about ways to describe things. Or, put another way, most all data is not schema nor descriptions of conceptual relationships, but making records available, with attributes and their values used to describe those records. Where is a business located? What political party does a politician belong to? How tall are you? What is the population of Hungary?

These are simple constructs with simple key-value pair ways to describe and convey them. This very simplicity is one reason why naïve data structs or simple data models like JSON or XML have proven so popular [4]. It is one of the reasons why the so-called NoSQL databases have also been growing in popularity. What we have are lots of atomic facts, located everywhere, and representable with very simple key-value structures.

While having such information available in linked data form makes it easier for agents to consume it, that extra publishing burden is by no means necessary. There are plenty of ways to consume that data -- without loss of information -- in non-linked data form. In fact, that is how the overwhelming percentage of such data is expressed today. This non-linked data is also often easy to understand.

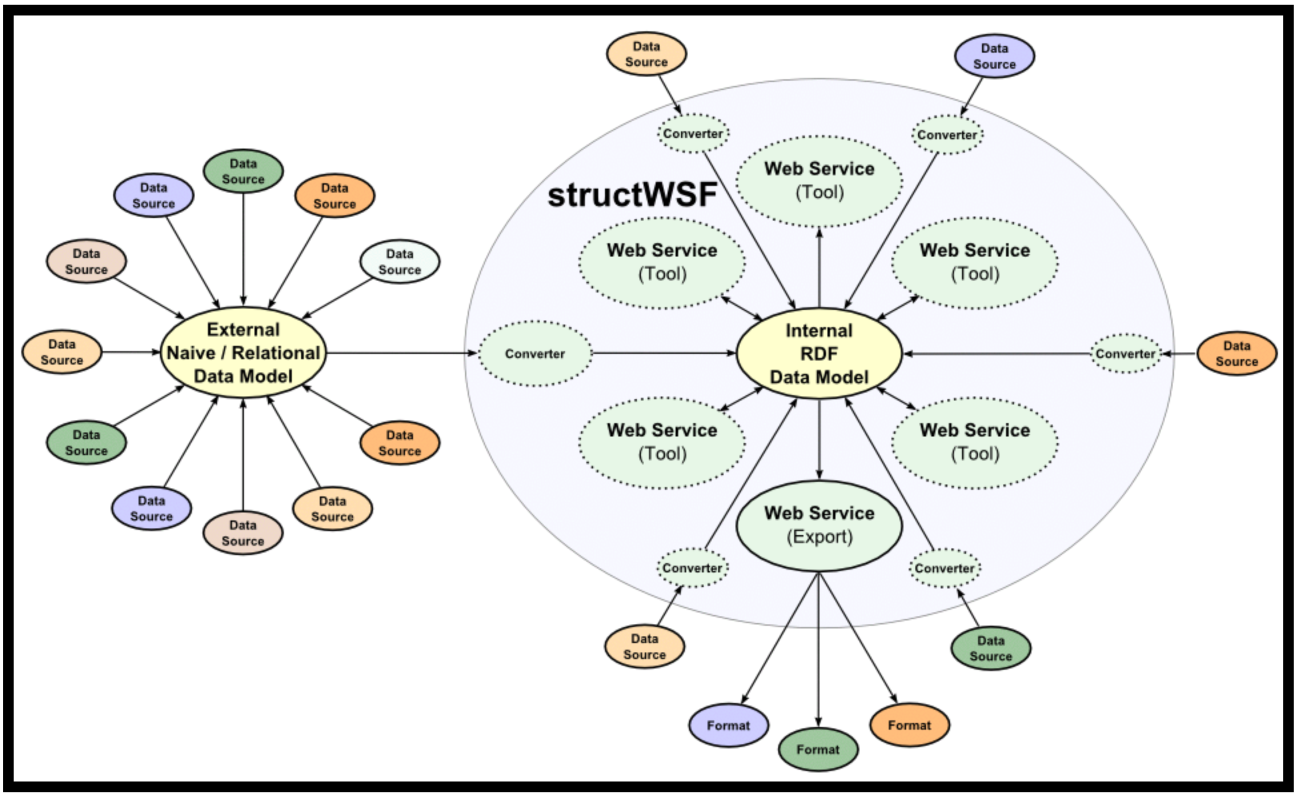
What ***is*** important is that the data be available electronically with a description of what the records contain. But that hurdle is met in many, many different ways and from many, many sources without any reference whatsoever to linked data. I submit that any form of desirable data available on the Web can be readily consumed without recourse to linked data principles.

## Observation #3: An Interoperable Data Model Does Not Require a Single Transmittal Format

The real advantage of RDF is the simplicity of its data model, which can be extended and augmented to express vocabularies and relationships of any nature. As I have stated before, that makes RDF like a *universal solvent* for any extant data structure, form or schema.

What I find perplexing, however, is how this strength somehow gets translated into a parallel belief that such a flexible data model is also the best means for ***transmitting*** data. As noted, most transmitted data can be represented through simple key-value pairs. Sure, at some point one needs to model the structural assumptions of the data model from the supplying publisher, but that complexity need not burden the actual transmitted form. So long as schema can be captured and modeled at the receiving end, data record transmittal can be made quite a bit simpler.

Under this mindset RDF provides the internal (canonical) data model. Prior to that, format and other converters can be used to consume the source data in its native form. A generalized representation for how this can work is shown in this diagram using Structured Dynamics' structWSF Web services framework middleware as the mediating layer:



Of course, if the source data is already in linked data form with understood concepts, relationships and semantics, much of this conversion overhead can be bypassed. If available, that is a good thing.

But it is not a required or necessary thing. Insistence on publishing data in certain forms suffers from the same narrowness as cultural or religious zealotry. Why certain publishers or authors prefer different data formats has a diversity of answers. Reasons can range from what is tried and familiar to available toolsets or even what is trendy, as one might argue linked data is in some circles today.There are literally scores of off-the-shelf "RDFizers" for converting native and simple data structs into RDF form. New converters are readily written.

Adaptive systems, by definition, do not require wholesale changes to existing practices and do not require effort where none is warranted. By posing the challenge as a "chicken-and-egg" one where publishers themselves must undertake a change in their existing practices to conform, or else they fail the "linked data threshold", advocates are ensuring failure. There is plenty of useful structured data to consume already.

Accessible structured data, properly characterized (see below), should be our root interest; not whether that data has been published as linked data *per se*.

## Observation #4: A Technique Can Not Carry the Burden of Usefulness or Interoperability

Linked data is nothing more than some techniques for publishing Web-accessible data using the RDF data model. Some have tried to use the concept of linked data as a replacement for the idea of the semantic Web, and some have recently tried to re-define linked data as not requiring RDF [5]. Yet the real issue with all of these attempts -- correct or not, and a fact of linked data since first formulated by Tim Berners-Lee -- is that a technique alone can not carry the burden of usefulness or interoperability.

Despite billions of triples now available, we in fact see little actual use or consumption of linked data, except in the life science domain. Indeed, a new workshop by the research community called COLD (Consuming Linked Data) has been set up for the upcoming ISWC conference to look into the very reasons why this lack of usage may be occurring [6].

It will be interesting to monitor what comes out of that workshop, but I have my own views as to what might be going on here. A number of factors, applicable frankly to any data, must be layered on top of linked data techniques in order for it to be useful:

- Context and coherence (see below)
- Curation and quality control (where provenance is used as the proxy), and
- Up-to-date and timely.

These requirements apply to any data ranging from Census CSV files to Google search results. But because relationships can also be more readily asserted with linked data, these requirements are even greater for it.

It is not surprising that the life sciences have seen more uptake of linked data. That community has keen experience with curation, and the quality and linkages asserted there are much superior to other areas of linked data [7].

In other linked data areas, it is really in limited pockets such as FactForge from Ontotext or curated forms of Wikipedia by the likes of Freebase that we see the most use and uptake. There is no substitute for consistency and quality control.

It is really in this area of "publish it and they will come" that we see one of the threads of parochialism in the linked data community. You can publish it and they still will ***not*** come. And, like any data, they will not come because the quality is poor or the linkages are wrong.

As a technique for making data available, linked data is thus nothing more than a foot soldier in the campaign to make information meaningful. Elevating it above its pay grade sets the wrong target and causes us to lose focus for what is really important.

## Observation #5: 50% of Linked Data is Missing (that is, the Linking part)

There is another strange phenomenon in the linked data movement: the almost total disregard for the linking part. Sure data is getting published as triples with dereferencable URIs, but where are the links?

At most, what we are seeing is `owl:sameAs` assertions and a few others [8]. Not only does this miss the whole point of linked data, but one can question whether equivalence assertions are correct in many instances [9].

For a couple of years now I have been arguing that the central gap in linked data has been the absence of context and coherence. By ***context*** I mean the use of reference structures to help place and frame what content is about. By ***coherence*** I mean that those contextual references make internal and logical sense, that they represent a consistent world view. Both require a richer use of links to concepts and subjects describing the semantics of the content.

It is precisely through these kinds of links that data from disparate sources and with different frames of reference can be meaningfully related to other data. This is the essence of the semantic Web and the purported purpose of linked data. And it is exactly these areas in which linked data is presently found most lacking.

Of course, these questions are not the sole challenge of linked data. They are the essential challenge in any attempt to connect or interoperate structured data within information systems. So, while linked data is ostensibly designed from the get-go to fulfill these aims, any data that can find meaning outside of its native silo must also be placed into ***context*** in a ***coherent*** manner. The unique disappointment for much linked data is its failure to provide these contexts despite its design.

## Observation #6: Pluralism is a Reality; Embrace It

Yet, having said all of this, Structured Dynamics is still committed to linked data. We present our information as such, and provide great tools for producing and consuming it. We have made it one of the seven foundations to our technology stack and methodology.

But we live in a pluralistic data world. There are reasons and roles for the multitude of popular structured data formats that presently exist. This inherent diversity is a fact in any real-world data context. Thus, we have not met a form of structured data that we didn't like, especially if it is accompanied with metadata that puts the data into coherent context. 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. irON clearly shows that entities can be usefully described and consumed in either RDF or non-RDF serialized forms.

Attitudes that dismiss non-linked data forms or arrogantly insist that publishers adhere to linked data practices are anything but pluralistic. They are parochial and short-sighted and are contributing, in part, to keeping the semantic Web from going mainstream.

Adoption requires simplicity. The simplest way to encourage the greater interoperability of data is to leverage existing assets in their native form, with encouragement for minor enhancements to add descriptive metadata for what the content is about. Embracing such an ecumenical attitude makes all publishers potentially valuable contributors to a better information future. It will also nearly instantaneously widen the tools base available for the common objective of interoperability.

## Parochialism and Root Cause Analysis

Linked data is a good thing, but not an ultimate thing. By making linked data an objective in itself we unduly raise publishing thresholds; we set our sights below the real problem to be solved; and we risk diluting the understanding of RDF from its natural role as a flexible and adaptive data model.

Paradoxically, too much parochial insistence on linked data may undercut its adoption and the realization of the overall semantic objective.

Root cause analysis for what it takes to achieve ***meaningful, interoperable information*** suggests that describing source content in terms of what it *is about* is the pivotal factor. Moreover, those contexts should be shared to aid interoperability. Whichever organizations do an excellent job of providing context and coherent linkages will be the go-to ones for data consumers. As we have seen to date, merely publishing linked data triples does not meet this test.

I have heard some state that first you celebrate linked data and its growing quantity, and then hope that the quality improves. This sentiment holds if indeed the community moves on to the questions of quality and relevance. The time for that transition is now. And, oh, by the way, as long as we are broadening our horizons, let's also celebrate properly characterized structured data no matter what its form. Pluralism is part of the tao to the meaning of information.

---

[1] See, for example, J.A. Hendler, 2008. "Web 3.0: Chicken Farms on the Semantic Web," *Computer*, January 2008, pp. 106-108. See http://www.comp.leeds.ac.uk/webscience/talks/hendler_web_3.pdf. While I can buy Hendler's arguments about commercial tool vendors holding off major investments until the market is sizable, I think we can also see via listings like Sweet Tools that a lack of tools is not in itself limiting.

[2] An earlier treatment of this subject from a different perspective is M.K. Bergman, 2010. "The Bipolar Disorder of Linked Data," *AI3:::Adaptive Information* blog, April 28, 2010.

[3] So far only prefixes for units up to 10^24 ("yotta") have names; for 10^27, a student campaign on Facebook is proposing "hellabyte" (North California slang for "a whole lot of") to get adopted by science bodies. See http://scitech.blogs.cnn.com/2010/03/04/hella-proposal-facebook/.

[4] One of more popular posts on this blog has been, M.K. Bergman, 2009. "'Structs': Naïve Data Formats and the ABox," *AI3:::Adaptive Information* blog, January 22, 2009.

[5] See, for example, the recent history on the linked data entry on Wikipedia or the assertions by Kingsley Idehen regarding entity attribute values (EAV) (see, for example, this blog post.)

[6] See further the 1st International Workshop on Consuming Linked Data (COLD 2010), at the 9th International Semantic Web Conference (ISWC 2010), November 8, 2010, Shanghai, China.

[7] For example, in the early years of GenBank, some claimed that annotations of gene sequences due to things like BLAST analyses may have had as high as 30% to 70% error rates due to propagation of initially mislabeled sequences. In part, the whole field of bioinformatics was formed to deal with issues of data quality and curation (in addition to analytics).

[8] 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.

[9] Harry Halpin and Patrick J. Hayes, 2010. "When owl:sameAs isn’t the Same: An Analysis of Identity Links on the Semantic Web," presented at *LDOW 2010*, April 27th, 2010, Raleigh, North Carolina. See http://events.linkeddata.org/ldow2010/papers/ldow2010_paper09.pdf.

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