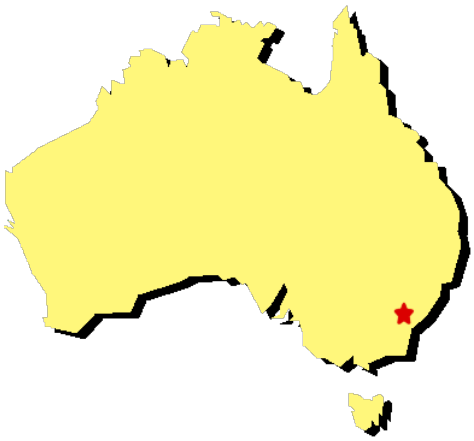
Pragmatic Approaches to the Semantic Web

by Mike Bergman - Monday, April 30, 2012

http://www.mkbergman.com/1006/pragmatic-approaches-to-the-semantic-web/



Linked Data is Sometimes a Useful Technique, but is an Inadequate Focus

While in Australia on other business, I had the great fortune to be invited by Adam Bell of the Australian War Memorial to be the featured speaker at the Canberra Semantic Web Meetup on April 23. The talk was held within the impressive BAE Systems Theatre of the Memorial and was very well attended. My talk was preceded by an excellent introduction to the semantic Web by David Ratcliffe and Armin Haller of CSIRO. They have kindly provided their useful slides online.

Many of the attendees came from the perspective of libraries, archives or museums. They naturally had an interest in the linked data activities in this area, a growing initiative that is now known under the acronym of LOD-LAM. Though I have been an advocate of linked data going back to 2006, one of my main theses was that linked data was an inadequate focus to achieve interoperability. The key emphases of my talk were that the pragmatic contributions of semantic technologies reside more in mindsets, information models and architectures than in 'linked data' as currently practiced.

Disappointments and Successes

The semantic Web and its most recent branding of linked data has antecedents going back to 1945 via Vannevar Bush's memex and Ted Nelson's hypertext of the early 1960s. The most powerful portrayal of the potential of the semantic Web comes in Douglas Adams' 1990 Hyperland special for the BBC, a full decade before Tim Berners-Lee and colleagues first coined the term 'semantic web' [1]. The Hyperland vision of obsequious intelligent agents doing our very bidding has, of course, not been fully realized. The lack of visible uptake of this full vision has caused some proponents to back away from the idea of the semantic Web. Linked data, in fact, was a term coined by Berners-Lee himself, arguably in part to re-brand the idea and to focus on a more immediate, achievable vision. In its first formulation linked data emphasized the RDF (Resource Description Framework) data model, though others, notably Kingsley Idehen, have attempted to put forward a revisionist definition of linked data that includes any form of

structured data involving entity attribute values (EAV).

No matter how expressed, the idea behind all of these various terms has in essence been to make meaningful connections, to provide the frameworks for interoperability. Interoperability means getting disparate sources of data to relate to each other, as a means of moving from data to information. Interoperability requires that source and receiver share a vocabulary about what things mean, as well as shared understandings about the associations or degree of relationship between the items being linked.

The current concept of linked data attempts to place these burdens mostly on the way data is published. While apparently "simpler" than earlier versions of the semantic Web (since linked data de-emphasizes shared vocabularies and nuanced associations), linked data places onerous burdens on how publishers express their data. Though many in the advocacy community point to the "billions" of RDF triples expressed as a success, actual consumers of linked data are rare. I know of no meaningful application or example where the consumption of linked data is an essential component.

However, there are a few areas of success in linked data. DBpedia, Freebase (now owned by Google), and GeoNames have been notable in providing identifiers (URIs) for common concepts, things, entities and places. There has also been success in the biomedical community with linked data.

Meanwhile, other aspects of the semantic Web have also shown success, but been quite hidden. Apple's spoken Siri service is driven by an ontological back-end; schema.org is beginning to provide shared ways for tagging key entities and concepts, as promoted by the leading search engines of Google, Bing, Yahoo! and Yandex; Bing itself has been improved as a search service by the incorporation of the semantic search technologies of its earlier Powerset acquisition; and Google is further showing how NLP (natural language processing) techniques can be used to extract meaningful structure for characterizing entities in search results and in search completion and machine language translation. These services are here today and widely used. All operate in the background.

What Lessons Can We Derive?

These failures and successes help provide some pragmatic lessons going forward.

While I disagree with Kingsley's revisionist approach to re-defining linked data, I very much agree with his underlying premise: effective data exchange does not require RDF. Most instance records are already expressed as simple entity-value pairs, and any data transfer serialization -- from key-value pairs to JSON to CSV spreadsheets -- can be readily transformed to RDF.

Semantic technologies are fundamentally about knowledge representation, not data transfer.

This understanding is important because the fundamental contribution of RDF is not as a data exchange format, but as a foundational data model. The simple triple model of RDF can easily express the information assertions in any form of content, from completely unstructured text (after information extraction or metadata characterization) to the most structured data sources. Triples can themselves be

built up into complete languages (such as OWL) that also capture the expressiveness necessary to represent any extant data or information schema [2].

The ability of RDF to capture any form of data or any existing schema makes it a "universal solvent" for information. This means that the real role of RDF is as a canonical data model at the core of the entire information architecture. Linked data, with its emphasis on data publishing and exchange, gets this focus exactly wrong. Linked data emphasizes RDF at the wrong end of the telescope.

The idea of common schema and representations is at the core of the semantic Web successes that do exist. In fact, when we look at Siri, emerging search, or some of the other successes noted above, we see that their semantic technology components are quite hidden. Successful semantics tend to work in the background, not in the foreground in terms of how data is either published or consumed. Semantic technologies are fundamentally about knowledge representation, not data transfer.

Where linked data is being consumed, it is within communities such as the life sciences where much work has gone into deriving shared vocabularies and semantics for linking and mapping data. These bases for community sharing express themselves as ontologies, which are really just formalized understandings of these shared languages in the applicable domain (life sciences, in this case). In these cases, curation and community processes for deriving shared languages are much more important to emphasize than how data gets exposed and published.

Linked data as presently advocated has the wrong focus. The techniques of publishing data and de-referencing URIs are given prominence over data quality, meaningful linkages (witness the appalling misuse of owl:sameAs [3]), and shared vocabularies. These are the reasons we see little meaningful consumption of linked data. It is also the reason that the much touted FYN ("follow your nose") plays no meaningful information role today other than a somewhat amusing diversion.

Shifting the Focus

In our own applications Structured Dynamics promotes seven pillars to pragmatic semantic technologies [4]. Linked data is one of those pillars, because where the other foundations are in place, including shared understandings, linked data is the most efficient data transfer format. But, as noted, linked data alone is insufficient.

Linked data is thus the wrong starting focus for new communities and users wishing to gain the advantages of interoperability. The benefits of interoperability must first obtain from a core (or canonical) data model -- RDF -- that is able to capture any extant data or schema. As these external representations get boiled down to a canonical form, there must be shared understandings and vocabularies to capture the meaning in this information. This puts community involvement and processes at the forefront of the semantic enterprise. Only after the community has derived these shared understandings should linked data be considered as the most efficient way to interchange data amongst the community members.

Identifying and solving the "wrong" problems is a recipe for disappointment. The challenges of the semantic Web are not in branding or messaging. The challenges of the semantic enterprise and Web reside more in mindsets, approaches and architecture. Linked data is merely a technique that contributes little -- perhaps worse by providing the wrong focus -- to solving the fundamental issue of information

interoperability.

Once this focus shifts, a number of new insights emerge. Structure is good in any form; arguments over serializations or data formats are silly and divert focus. The role of semantic technologies is likely to be a more hidden one, to reside in the background as current successes are now showing us. Building communities with trusted provenance and shared vocabularies (ontologies) are the essential starting points. Embracing and learning about NLP will be important to include the 80% of content currently in unstructured text and disambiguating reference conflicts. Ultimate users, subject matter experts and librarians are much more important contributors to this process than developers or computer scientists. We largely now have the necessary specifications and technologies in place; it is time for content and semantic reconciliation to guide the process.

It is great that the abiding interest in interoperability is leading to the creation of more and more communities, such as LOD-LAM, forming around the idea of linked data. What is important moving forward is to use these interests as springboards, and not boxes, for exploring the breadth of available semantic technologies.

For More on the Talk

Below is a link to my slides used in Canberra:

Pragmatic Approaches to the Semantic Web

View more presentations from Mike Bergman.

Also, as mentioned, the intro slides are online, a video recording of the presentations is also available, and some other blog postings occasioned by the talks are also online.

---

[1] Tim Berners-Lee, James Hendler and Ora Lassila, 2001. "The Semantic Web". *Scientific American Magazine*; see http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

[2] See further, M.K. Bergman, 2009. "Advantages and Myths of RDF," AI3:::Adaptive Innovation blog, April 8, 2009. See http://www.mkbergman.com/483/advantages-and-myths-of-rdf/.

[3] See, among many, M.K. Bergman, 2010. "Practical P-P-P-Problems with Linked Data," AI3:::Adaptive Innovation blog, October 4, 2010. See http://www.mkbergman.com/917/practical-p-p-p-problems-with-linked-data/.

[4] M.K. Bergman, 2010. "Seven Pillars of the Open Semantic Enterprise," AI3:::Adaptive Innovation blog, January 12, 2010. See http://www.mkbergman.com/859/seven-pillars-of-the-open-semantic-enterprise/.
