A Decade in the Trenches of the Semantic Web

by Mike Bergman - Wednesday, July 16, 2014

http://www.mkbergman.com/1771/a-decade-in-the-trenches-of-the-semantic-web/

Are We Losing the War? Was it Even the Right One?

Cinemaphiles will readily recognize Akira Kurosawa's Rashomon film of 1951. And, in the 1960s, one of the most popular book series was Lawrence Durrell's The Alexandria Quartet. Both, each in its own way, tried to get at the question of what is truth by telling the same story from the perspective of different protagonists. Whether you saw this movie or read these books you know the punchline: the truth was very different depending on the point of view and experience -- including self-interest and delusion -- of each protagonist. All of us recognize this phenomenon of the blind men's view of the elephant.

I have been making my living and working full time on the semantic Web and semantic technologies now for a full decade. So has my partner at Structured Dynamics, Fred Giasson. Others have certainly worked longer in this field. The original semantic Web article appeared in *Scientific American* in 2000 [1], and the foundational Resource Description Framework data model dates from 1999. Fred and I have our own views of what has gone on in the trenches of the semantic Web over this period. We thought a decade was a good point to look back, share what we've experienced, and discover where to point our next offensive thrusts.

What Has Gone Well?

The vision of the semantic Web in the *Scientific American* article painted a picture of globally interconnected data leveraged by agents or bots designed to make our lives easier and more automated. However, by the time that I got directly involved, nearly five years after standards first started to be published, Tim Berners-Lee and many leading proponents of RDF were beginning to shift focus to linked data. The agents, and automation, and ontologies of the initial vision were being downplayed in favor of effective means to publish and consume data based on RDF. In many ways, linked data resembled a re-branding.

This break had been coming for a while, memorably captured by a 2008 ISWC session led by Peter F.

Patel-Schneider [2]. This internal division of viewpoint likely caused effort to be split that would have been better spent in proselytizing and improving tools. It also diverted somewhat into internal squabbles. While many others have pointed to a tactical mistake of using an XML serialization for early versions of RDF as a key factor is slowing initial adoption, a factor I agree was at play, my own suspicion is that the philosophical split taking place in the community was the heavier burden.

Whatever the cause, many of the hopes of the heady days of the initial vision have not been obtained over the past fifteen years, though there have been notable successes.

The biomedical community has been the shining exemplar for data interoperability across an entire discipline, with earth sciences, ecology and other science-based domains also showing interoperability success [3]. Families of ontologies accompanied by tooling and best practices have characterized many of these efforts. Sadly, though, most other domains have not followed suit, and commercial interoperability is nearly non-existent.

Most all of the remaining success has resided in single-institution data integration and knowledge representation initiatives. IBM's Watson and Apple's Siri are two amazing capabilities run and managed by single institutions, as is Google's Knowledge Graph. Also, some individual commercial and government enterprises, willing to pay support to semantic technology experts, have shown success in data integration, using RDF, SKOS and OWL.

We have seen the close kinship between natural language, text, and Q & A with the semantic Web, also demonstrated by Siri and more recent offshoots. We have seen a trend toward pairing great-performing open source text engines, notably Solr, with RDF and triple stores. Recommendation systems have shown some success. Linked data publishing has also had some notable examples, including the first of the lot, DBpedia, with certain institutional publishers (such as the Library of Congress, Eurostat, The Getty, Europeana, OpenGLAM [galleries, archives, libraries, and museums]) showing leadership and the commitment of significant vocabularies to linked data form.

On the standards front, early experience led to new and better versions of the SPARQL query language (SPARQL 1.1 was greatly improved in the last decade and appears to be one capability that sells triple stores), RDF 1.1 and OWL 2. Certain open source tools have become prominent, including Protégé, Virtuoso (open source) and Jena (among unnamed others, of course). At least in the early part of this history, tool development was rapid and flourishing, though the innovation pace has dropped substantially according to my tracking database Sweet Tools.

What Has Disappointed?

My biggest disappointments have been, first, the complete lack of distributed data interoperability, and, second, the lack or inability of commercial enterprises to embrace and adopt semantic technologies on their own. The near absence of discussion about instance records and their attributes helps frame the current maturity of the semantic Web. Namely, it has yet to crack the real nuts of data integration and interoperability across organizations. Again, with the exception of the biomedical community, neither in the linked data realm nor in the broader semantic Web, can we point to information based on semantic Web principles being widely shared between systems and organizations.

Some in the linked data community have explicitly acknowledged this. The abstract for the upcoming COLD 2014 workshop, for example, states [4]:

. . . applications that consume Linked Data are not yet widespread. Reasons may include a lack of suitable methods for a number of open problems, including the seamless integration of Linked Data from multiple sources, dynamic discovery of available data and data sources, provenance and information quality assessment, application development environments, and appropriate end user interfaces.

We have written about many issues with linked data, ranging from the use of improper mapping predicates; to the difficulty in publishing; and to dereferencing URIs on the Web since they are sparse and not always properly implemented [5]. But ultimately, most linked data is just instance data that can be represented in simpler attribute-value form. By shunning a knowledge representation language (namely, OWL) at the processing end, we have put too much burden on what are really just instance records. Linked data does not get the balance of labor right. It ignores the reality that data consumers want actionable information over being able to click from data item to data item, with overall quality reduced to the lowest common denominator. If a publisher has the interest and capability to publish quality linked data, great! It should become part of the data ingest pool and the data becomes easy to consume. But to insist on linked data across the board creates unnecessary barriers. Linked data growth has not nearly kept pace with broader structured data growth on the Web [6].

At the enterprise level, the semantic technology stack is hard to grasp and understand for newcomers. RDF and OWL awareness and understanding are nearly nil in companies without prior semantic Web experience, or 99.9% of all companies. This is not a failure of the enterprises; it is the failure of us, the advocates and suppliers. While we (Structured Dynamics) have developed and continue to refine the turnkey Open Semantic Framework stack, and have spent more efforts than most in documenting and explicating its use, the systems are still too complicated. We combine complicated content management systems as user front-ends to a complicated semantic technology stack that needs to be driven by a complicated (to develop) ontology. And we think we are doing some of the best technology transfer around!

Moreover, while these systems are good at integrating concepts and schema, they are virtually silent on the question of actual data integration. It is shocking to say, but the semantic Web has no vocabularies or tools sufficient to enable data items for the same entity from two different datasets to be combined or reconciled [7]. These issues can be solved within the individual enterprise, but again the system breaks when distributed interoperability is the desire. General Web-based inconsistencies, such as in HTML coding or mime types, impose hurdles on distributed interoperability. These are some of the reasons why we see the successes in the context (generally) of single institutions, as opposed to anything that is truly yet Web-wide.

These points, as is often the case with software-oriented technologies, come down to a disappointing state of tooling. Markets drive developer interest, and market share has been disappointing; thus, fewer tools. Tool interest comes from commercial engagements, and not generally grants, the major source of semantic Web funding, particularly in the European Union. Pragmatic tools that solve real problems in user adoption are rarely a sufficient basis for getting a Ph.D.

The weaknesses in tooling extend from basic installation, to configuration, unit and integrated tests, data conversion and lifting, and, especially, all things ontology. Weaknesses in ontology tooling include (critically) mapping, consistency and coherency checking, authoring, managing, version control, re-factoring, optimization, and workflows. All of these issues are solvable; they are standard software challenges. But it is hard to conquer markets largely with the wrong army pursuing the wrong objectives in response to the wrong incentives.

Yet, despite the weaknesses in tooling, we believe we have been fairly effective in transferring technology to our clients. It takes more documentation and more training and, often, accompanying tool development or improvement in the workflow areas critical to the project. But clients need to be told this as well. In these still early stages, successful clients are going to have to expend more staff effort. With reasonable commitment, it is demonstrable that an enterprise can take over and manage a large-scale semantic engagement on its own. Still, for semantic technologies to have greater market penetration, it will be necessary to lower those commitments.

How Has the Environment Changed?

Of course, over the period of this history, the environment as a whole has changed markedly. The Web today is almost unrecognizable from the Web of 15 years ago. If one assumes that Web technologies tend to have a five year or so period of turnover, we have gone through at least two to three generations of change on the Web since the initial vision for the semantic Web.

The most systemic changes in this period have been cloud computing and the adoption of the smartphone. These, plus the network of workstations approach to data centers, have radically changed what is desirable in a large-scale, distributed architecture. APIs have become RESTful and database infrastructures have become flatter and more distributed. These architectures and their supporting infrastructure -- such as virtual servers, MapReduce variants, and many applications -- have in turn opened the door to performant management of large volumes of flat (key-value or graph) data, or big data .

On the Web side, JavaScript, just a few years older than the semantic Web, is now dominant in Web pages and taking on server-side roles (such as through Node.js). In turn, JSON has now grown in popularity as a form of data representation and transfer and is being adopted to the semantic Web (along with codifying CSV). Mobile, too, affects the Web side because of the need for multiple-platform deployments, touchscreen use, and different user interface paradigms and layout designs. The app ecosystem around smartphones has become a huge source for change and innovation.

Extremely germane to the semantic Web -- indeed, overall, for artificial intelligence -- has been the occurrence of knowledge-based AI (KBAI). The marrying of electronic Web knowledge bases -- such as Wikipedia or internal ones like the Google search index or its Knowledge Graph -- with improvements in machine-learning algorithms is systematically mowing down what used to be called the Grand Challenges of computing. Sensors are also now entering the picture, from our phones to our homes and our cars, that exposes the higher-order requirement for data integration combined with semantics. NLP kits have improved in terms of accuracy and execution speed; many semantic tasks such as tagging or categorizing or questioning already perform at acceptable levels for most projects.

On the tooling side, nearly all building blocks for what needs to be done next are available in open source, with some platform areas quite functional (including OSF, of course). We have also been successful in finding clients that agree to open source the development work we do for them, since they are benefiting from the open source development that went on before them.

What Did We Set Out to Achieve?

When Structured Dynamics entered the picture, there were already many tools available and core languages had been released. Our view of the world at that time led us to adopt two priorities for what we thought might be a five year or so plan. We have achieved the objectives we set for ourselves then, though it has taken us a couple of years longer to realize.

One priority was to develop a reference structure for concepts to serve as a "grounding" basis for relating datasets, vocabularies, schema, taxonomies, or ontologies. We achieved this with our first commercial release (v 1.00) of UMBEL in February 2011. Subsequent to that we have progressed to v 1.05. In the coming months we will see two further major updates that have been under active effort for about eight months.

The other priority was to create a turnkey foundation for a semantic enterprise. This, too, has been achieved, with many more releases. The Open Semantic Framework (OSF) is now in version 3.00, backed by a 500-article training documentation and technical wiki. Support tooling now includes automated installation, testing, and data transfer and synchronization.

Because our corporate objectives were largely achieved it was time to look at lessons learned and set new directions. This article, in part, is a result of that process.

How Did Our Priorities Evolve Over the Decade?

I thought it would be helpful to use the content of this **AI3** blog to track how concerns and priorities changed for me and Structured Dynamics over this history. Since I started my blog quite soon after my entry into the semantic Web, the record of my perspectives was conterminous and rather complete.

The fifty articles below trace my evolution in knowledge and skills, as well as a progression from structured data to the semantic Web. These 50 articles represent about 11% of all articles in my chronological archive; they were selected as being the most germane to the question of evolution of the semantic Web.

After early ramp up, most of the formative discussion below occurred in the early years. Posts have declined most recently as implementation has taken over. Note most of the links below have PDFs available from their main pages.

2014

- Innovation, Information, Growth and Wealth - information fuels innovation that creates wealth
- Spring Dawns on Artificial Intelligence - massive trends are waking artificial intelligence (AI)

from its dark winter

2013

- Seven Arguments for Semantic Technologies - a re-cap and summary of prior writings

2012

- The Age of the Graph - the ubiquitous and fundamental roles of graph structures
- The Rationale for Semantic Technologies - most refined arguments to date
- What is Structure? - structure is information, and information is structure
- The Trouble with Memes - the role of Shannon's information theory to adaptive information
- Give Me a Sign: What Do Things Mean on the Semantic Web? - nature of meaning and representation; influence of Charles S Peirce

2011

- Making the Argument for Semantic Technologies - five unique advantages for the enterprise
- In the Midst of an Evolutionary Explosion - artificial intelligence (AI) finally begins to flex its muscles
- Leveraging Intangible Assets Using Semantic Technologies - in a knowledge economy, the value of intangible assets exceeds tangible ones
- Democratizing Information with Semantics - putting the IT function into the hands of users, the knowledge workers
- Ontology-Driven Apps Using Generic Applications - the technology is here to stand software engineering on its head
- Seeking a Semantic Web Sweet Spot - making the argument for reference structures
- Making Connections Real - role of knowledge bases in guiding connections
- Declining IT Innovation in the Enterprise - innovation is shifting to the consumer sector

2010

- What is a Reference Concept? - information interoperability requires some fixed reference points
- The Nature of Connectedness on the Web - the reality is most connections are proximate
- A Reference Guide to Ontology Best Practices - as stated
- I Have Yet to Metadata I Didn't Like - the real world issue is not how to publish data, but how to consume and curate it
- An Executive Intro to Ontologies - first recommended resource for learning about ontologies
- 'Pay as You Benefit': A New Enterprise IT Strategy - using incremental, low-risk and open approaches to adopting semantics
- Changing IT for Good - how to transition the enterprise to semantic technologies
- Seven Pillars of the Open Semantic Enterprise - fundamental overview of the semantic enterprise; one of my most cited articles

2009

- The Open World Assumption: Elephant in the Room - the fundamental importance of open world reasoning to knowledge applications
- Ontology-driven Applications Using Adaptive Ontologies - using ontologies as the central governing structures for semantic technologies
- When Linked Data Rules Fail - questioning the limits of linked data as often practiced
- The Law of Linked Data - quantifies the benefits from interconnecting data
- Fresh Perspectives on the Semantic Enterprise - first description of how semantic technologies can fit within the enterprise
- Structure the World - an integrative view of how native data forms can integrate into the semantic Web
- structWSF: A Framework for Collaboration Networks - how architectural design can promote collaboration and distributed semantic Web
- Advantages and Myths of RDF - my definitive piece on RDF (Resource Description Framework)
- Making Linked Data Reasonable using Description Logics, Part 3 - generalizing how native data forms can interact with the semantic Web
- Making Linked Data Reasonable using Description Logics, Part 2 - how to split work between the TBox, ABox and specialty services
- Back to the Future with Description Logics - first description of the importance of keeping schema (TBox) separate from instance records (ABox)

2008

- Thinking Inside the Box with Description Logics - beginning to explicate the logic underneath W3C semantic technologies
- WOA: A New Enterprise Partner for Linked Data - the importance of architecture for emerging solutions
- When is Content Coherent? - the essential metric of 'coherence' to semantic vocabularies
- The Semantic Web and Industry Standards - early understandings of the Semantic Web

2007

- A Data Model of Web Data Models: Part I - good structural overview, now 7 years old !
- What is the Structured Web? - explication of the structured Web as an intermediate way point to the semantic Web
- Announcing UMBEL: A Lightweight Subject Structure for the Web - first announcement for the UMBEL reference concept structure
- Where are the Road Signs for the Structured Web? - first identification that the semantic Web is missing a system of reference concepts
- Structure Paves the Way to the Semantic Web - invited guest editorial in IEEE Intelligent Systems
- There's Not Yet Enough Backbone - a suitable ***subject*** structure for organizing knowledge is needed
- Structurizing the Web with RDF - first rather comprehensive piece on the benefits of RDF data model
- Did You Blink? The Structured Web Just Arrived - April 2007; first external piece on DBpedia

2006

- Sources and Classification of Semantic Heterogeneities - still one of the best primers on heterogeneous data
- Climbing the Data Federation Pyramid - puts Internet and semantic Web into context

2005

- Open Source and the 'Business Ecosystem' - the importance of keystone influencers and partnerships to build technology ecosystems

The early years of this history were concentrated on gathering background information and getting educated. The release of DBpedia in 2007 showed how knowledge bases would become essential to the semantic Web. We also identified that a lack of shared reference concepts was making it difficult to "ground" different semantic Web datasets or schema to one another. Another key theme was the diversity of native data structures on the Web, but also how all of them could be readily represented in RDF.

By 2008 we began to study the logical underpinnings to the semantic Web as we were coming to understand how it should be practiced. We also began studying Web-oriented architectures as key design guidance going forward. These themes continued into 2009, though now informed by clients and applications, which was expanding our understanding of requirements (and, sometimes, shortcomings) in the enterprise marketplace. The importance of an open world approach to the basic open nature of knowledge management was cementing a clarity of the role and fit of semantic solutions in the overall informaton space. The general community shift to linked data was beginning to surface worries.

2010 marked a shift for us to become more of a popularizer of semantic technologies in the enterprise, useful to attract and inform prospects. The central role of ontologies as the guiding structures (either as codified knowledge structures or as instruction sets for the platform) for OSF opened realizations that generic functional software could be designed that can be re-used in most any knowledge domain by simply changing the data and ontologies guiding them. This increased our efforts in ontology tooling and training, now geared more to the knowledge worker. The importance of groundings for aligning schema and data caused us to work hard on UMBEL in 2011 to get it to a commercial release state.

All of these efforts were converging on design thoughts about the nature of information and how it is signified and communicated. The bases of an overall philosophy regarding our work emerged around the teachings of Charles S Peirce and Claude Shannon. Semantics and groundings were clearly essential to convey accurate messages. Simple forms, so long as they are correct, are always preferred over complex ones because message transmittal is more efficient and less subject to losses (inaccuracies). How these structures could be represented in graphs affirmed the structural correctness of the design approach. The now obvious re-awakening of artificial intelligence helps to put the semantic Web in context: a key subpart, but still a subset, of artificial intelligence. The percentage of formative articles directly related over these last couple of years to the semantic Web drops much, as the emphasis continues to shift to tech transfer.

What Else Did We Learn?

Not all lessons learned warranted an article on their own. So, we have also reflected on what other lessons we learned over this decade. The overall theme is: Simpler is better.

Distributed data interoperability across the Web is a fundamental weakness. There are no magic tricks to integrate data. Data mapping and integration will always require massaging. Each data integration activity needs its own solution. However, it can greatly be helped with ontologies and with better tooling.

In keeping with the lesson of grounding, a reference ontology for attributes is missing. It is needed as a bridge across disparate datasets describing similar entities or with different attributes for the same entities. It is also a means to reduce the pairwise combinatorial issue of integrating multiple datasets. And, whatever is done in the data integration area, an open world approach will be essential given the nature of knowledge information.

There is good design and best practice for distributed architectures. The larger these installations become, the more important it is to use a lightweight, loosely-coupled design. RESTful Web services and their interfaces are key. Simpler services with fewer functions can be designed to complement one another and increase throughput effectiveness.

Functional programming languages align well with the data and schema in knowledge management functions. Ontologies, as structures, also fit well with functional languages. The ability to create DSLs should continue to improve bringing the knowledge management function directly into the hands of its users, the knowledge workers.

In a broader sense, alluded to above, the semantic Web is but a set of concepts. There are multiple ways to use it. It can be leveraged without requiring "core" semantic Web tools such a triple stores. Solr can act as a semantic store because semantics, NLP and search are naturally married. But, the semantic Web, in turn, needs to become re-embedded in artificial intelligence, now backed by knowledge bases, which are themselves creatures of the semantic Web.

Design needs to move away from linked data or the semantic Web as the goals. The building blocks are there, though perhaps not yet combined or expressed well. The real improvements now to the overall knowledge function will result from knowledge bases, artificial intelligence, and the semantic Web working together. That is the next frontier.

Overall, we perhaps have been in the wrong war for the wrong reasons. Linked data is certainly not an end and mostly appears to represent work, rather than innovation. The semantic Web is no longer the right war, either, because improvements there will not come so much from arguing semantic languages and paradigms. Learning how to master distributed data integration will teach the semantic Web much, and coupling artificial intelligence with knowledge bases will do much to improve the most labor-intensive stumbling blocks in the knowledge management workflow: mappings and transformations. Further, these same bases will extend the reach into analytical and statistical realms.

The semantic Web has always been an infrastructure play to us. On that basis, it will be hard to ever judge market penetration or dominance. So, maybe in terms of a vision from 15 years ago the growth of the semantic Web has been disappointing. But, for Fred and me, we are finally seeing the landscape clearly

and in perspective, even if from a viewpoint that may be different from others'. From our vantage point, we are at the exciting cusp of a new, broader synthesis.

NOTE: This is Part I of a two-part series. Part II will appear shortly.

---

[1] Tim Berners-Lee, James Hendler, and Ora Lassila, “The Semantic Web,” in *Scientific American* 284(5): pp 34-43, 2001. See http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2.

[2] For those with a spare 90 minutes or so, you may also want to view this panel session and debate that took place on “An OWL 2 Far?” at ISWC ’08 in Karlsruhe, Germany, on October 28, 2008. The panel was chaired by Peter F. Patel-Schneider (Bell Labs, Alcathor) with the panel members of Stefan Decker (DERI Galway), Michel Dumontier (Carleton University), Tim Finin (University of Maryland) and Ian Horrocks (University of Oxford), with much audience participation. See http://videolectures.net/iswc08_panel_schneider_owl/

[3] Open Biomedical Ontologies (OBO) is an effort to create controlled vocabularies for shared use across different biological and medical domains. As of 2006, OBO formed part of the resources of the U.S. National Center for Biomedical Ontology (NCBO). As of the date of this article, there were 376 ontologies listed on the NCBO's BioOntology site. Both OBO and BioOntology provide tools and best practices.

[4] Fifth International Workshop on Consuming Linked Data (COLD 2014), co-located with the 13th International Semantic Web Conference (ISWC) in Riva del Garda, Italy, October 19-20.

[5] See http://www.mkbergman.com/category/linked-data/.

[6] See http://www.mkbergman.com/1713/fyn-v-ft/.

[7] See the thread on the W3C semantic web mailing list beginning at http://lists.w3.org/Archives/Public/semantic-web/2014Jul/0129.html.
