Big Structure: At The Nexus of Knowledge Bases, the Semantic Web and Artificial Intelligence

by Mike Bergman - Wednesday, July 23, 2014

http://www.mkbergman.com/1773/big-structure-at-the-nexus-of-knowledge-bases-the-semantic-web-and-artificial-intelligence/

Envisioning A New Adaptive Infrastructure for Data Interoperability

In Part I of this two-part series, Fred Giasson and I looked back over a decade of working within the semantic Web and found it partially successful but really the wrong question moving forward. The inadequacies of the semantic Web to date reside in its lack of attention to practical data interoperability across organizational or community boundaries. An emphasis on linked data has created an illusion that questions of data integration are being effectively addressed. They are not.

Linked data is hard to publish and not the only useful form for consuming data; linked data quality is often unreliable; the linking predicates for relating disparate data sources to one another may be inadequate or wrong; and, there are no reference groundings for relating data values across datasets. Neither the semantic Web nor linked data has developed the practices, tooling or experience to actually interoperate data across the Web. These criticisms are not meant to condemn linked data -- it is, after all, the early years. Where it is compliant and from authoritative information sources, linked data can be a gold standard in data publishing. But, linked data is neither necessary nor essential, and may even be a diversion if it sucks the air from the room for what is more broadly useful.

This table summarizes the state-of-art in the semantic Web for frameworks and guidance in how to interoperate data:

Category	Related Terms	Status in the Semantic Web
Classes	sets, concepts, topics, types, kinds	Mature, but broader scope coverage desirable; equivalent linkages between datasets often mis-

		applied; more realistic proximate linkages in flux, with no bases to reason over them
Instances	individuals, entities, members, records, things	Current basis for linked data; many linkage properties mis-applied
Relation Properties	relations, predicates	Equivalent linkages between datasets often mis-applied; more realistic proximate linkages in flux, with no bases to reason over them.
Descriptive Properties	attributes, descriptors	Save for a couple of minor exceptions, no basis for mapping attributes across datasets
Values	data	Basic QUDT ontologies could contribute here

We can relate the standard *subject - predicate - object* triple statement in RDF to this table, using the **Category** column. Classes and Instances relate to the *subjects*, Relation and Descriptive Properties relate to the *predicate*, and Values relate to the *object* [6] in an RDF triple. The concepts and class schema of different information sources (their "aboutness") can reasonably be made to interoperate. In terms of the description logics that underly the logic bases of W3C ontologies, the focus and early accomplishments of the semantic Web have been on this "terminological box" or T-Box [7]. Tooling to make the mappings more productive and means to test the coherence and completeness of the results still remain as priority efforts, but the conceptual basis and best practices have progressed pretty well.

In contrast, nearly lacking in focus and tooling has been the flip side of that description logics coin: the A-Box [7], or assertional and instance (data) level of the equation. Both the T-Box and A-Box are necessary to provide a knowledge base. Today, there are virtually no vocabularies, no tooling, no history, no best practices and no "grounding" for actual A-Box data integration within the semantic Web. Without such guidance, the semantic Web is silent on the questions of data interoperability. As David Karger explained in his keynote address at ISWC in 2013 [8], "we've got our heads in the clouds while people are stuck in the dirt."

Yet these are not fatal flaws of the semantic Web, nor are they permanent. Careful inspection of current circumstances, combined with purposeful action, suggests:

1. Data integration can be solved
2. Leveraging background knowledge is a key enabler
3. Interoperability requires reference structures, what we are calling *Big Structure*.

The Prism of Data Interoperability

Why do we keep pointing to the question of data interoperability? Consider these facts:

- 80% of all available information is in text or documents (unstructured)
- 40% of standard IT project expenses are devoted to data integration in one form or another, due to the manual effort needed for data migration and mapping
- Information volumes are now doubling in fewer than two years
- Other trends including smartphones and sensors are further accelerating information growth
- Effective business intelligence requires the use of quality, integrated data.

The abiding, costly, frustrating and energy-sucking demands of data integration have been a constant within enterprises for more than three decades. The same challenges reside for the Web. The Internet of Things will further demand better interoperability frameworks and guidelines. Current data integration tooling relies little upon semantics and no leading alternative is based principally around semantic approaches [9].

The data integration market is considered to include enterprise data integration and extract, transform and load (ETL) vendors. Gartner estimates tool sales for this market to be about $2 billion annually, with a growth rate faster than most IT areas [10]. But data integration also touches upon broader areas such as enterprise application integration (EAI), federated search and query, and master data management (MDM), among others. Given that data integration is also 40% of standard IT project costs, new approaches are needed to finally unblock the costly logjam of enterprise information integration. Most analysts see firms that are actively pursuing data integration innovations as forward-thinking and more competitive.

Data integration is combining information from multiple sources and providing users a uniform view of it. *Data interoperability* is being able to exchange and work upon (inter-operate) information across system and organizational boundaries. The ability to integrate data precedes the ability to interoperate it. For example, I may have three datasets of mammals that I want to consolidate and describe in similar terms with common units of measurement. That is an example of data integration. I may then want to relate this mammal knowledge base with a more general perspective of the animal kingdom. That is an example of data interoperability. Data integration usually occurs within a single organization or enterprise or institutional offering (as would be, say, Wikipedia). Data interoperability additionally needs to define meanings and communicate them in common ways across organizational, domain or community boundaries.

These are natural applications for the semantic Web. Why, then, has there not been more practical use of the semantic Web for these purposes?

That is an interesting question that we only partially addressed in Part I of this series. All aspects of data have semantics: what the data is about, what is its context, how it relates to other data, and what its values are and what they mean. The semantic Web is closely allied with natural language processing, an essential for bringing the 80% of unstructured data into the equation. Semantic Web ontologies are useful structures for how to relate real-world data into common, reference forms. The open world logic of the semantic Web is the right perspective for knowledge functions under the real-world conditions of constantly expanding information and understandings.

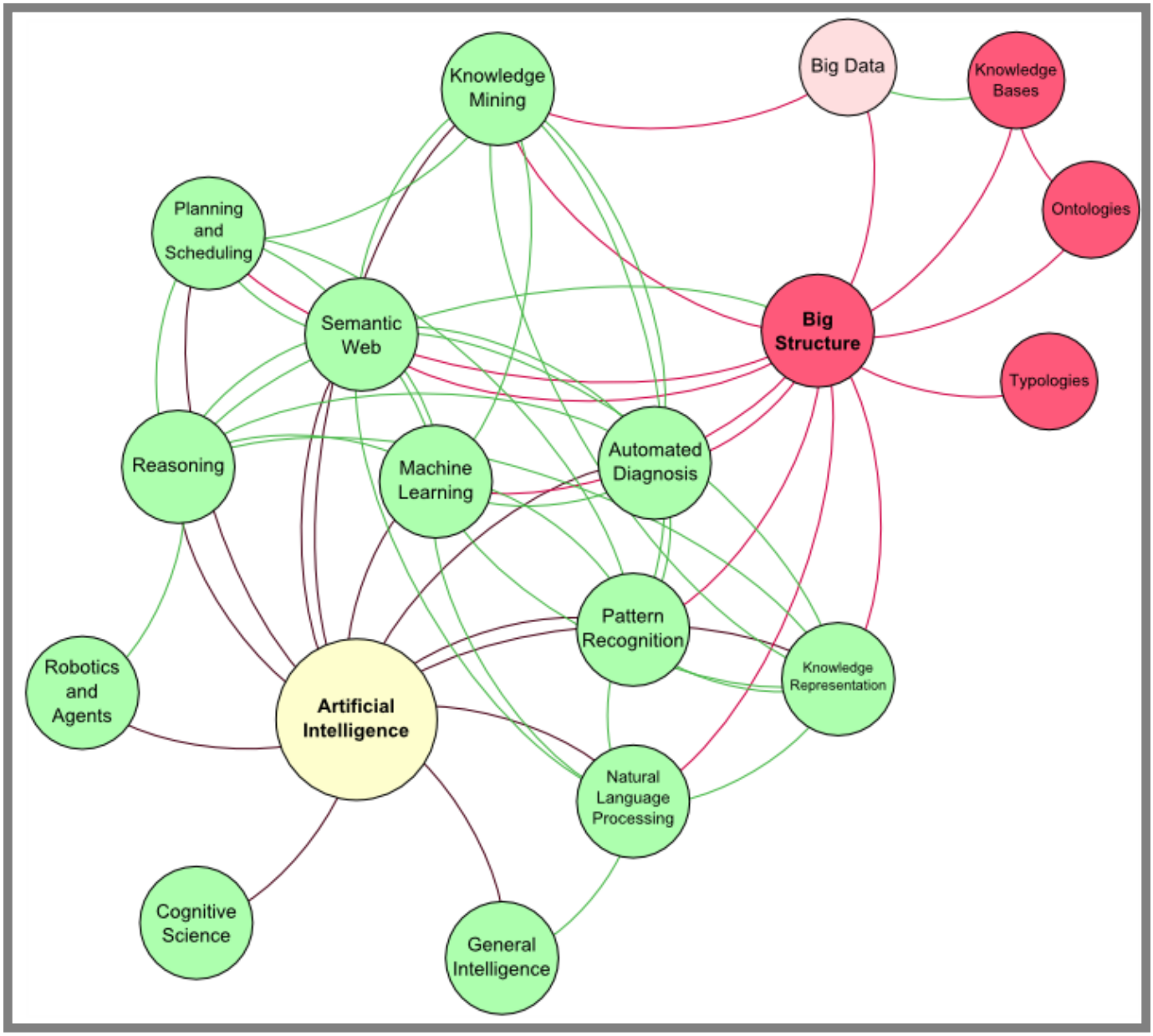
While these requirements suggest an integral role for the semantic Web, it is also clear that the semantic Web has not yet made these contributions. One explanation may be that semantic Web advocates, let

alone the linked data tribe, have not seen data integration -- as traditionally defined -- as their central remit. Another possibility is that trying to solve data interoperability through the primary lens of the semantic Web is the wrong focus. In any case, meeting the challenge of data interoperability clearly requires a much broader context.

Embedding Data Interoperability Into a Broader Context

The semantic Web, in our view, is properly understood as a sub-domain of artificial intelligence. Semantic technologies mesh smoothly with natural language tasks and objectives. But, as we noted in a recent review article, artificial intelligence is itself undergoing a renaissance [11]. These advances are coming about because of the use of knowledge-based AI (KBAI), which combines knowledge bases with machine learning and other AI approaches. Natural language and spoken interfaces combined with background knowledge and a few machine-language utilities are what underlie Apple's Siri, for example.

The realization that the semantic Web is useful but insufficient and that AI is benefitting from the leveraging of background knowledge and knowledge bases caused us to "decompose" the data-interoperability information space. Because artificial intelligence is a key player here, we also wanted to capture all of the main sub-domains of AI and their relationships to one another:


Knowledge Mining
Big Data
Knowledge Bases
Planning and Scheduling
Ontologies
Semantic Web
Big Structure
Typologies
Reasoning
Machine Learning
Automated Diagnosis
Pattern Recognition
Robotics and Agents
Artificial Intelligence
Knowledge Representation
Natural Language Processing
Cognitive Science
General Intelligence

Artificial Intelligence Domains

Two core observations emerge from standing back and looking at these questions. First, many of AI's main sub-domains have a role to play with respect to data integration and interoperability:

AI Domains Related to Data Interoperability

This places semantic Web technologies as a co-participant with natural language processing, knowledge mining, pattern recognizers, KR languages, reasoners, and machine learning as domains related to data interoperability.

And, second, generalizing the understanding of knowledge bases and other guiding structures in this space, such as ontologies, highlights the potential importance of *Big Structure*. Virtually every one of the domains displayed above would be aided by leveraging background knowledge.

Grounding Data Interoperability in Big Structure

As our previous AI review showed [11], reference knowledge bases -- Wikipedia in the forefront -- have been a tremendous boon to moving forward on many AI challenges. Our own experience with UMBEL has also shown how reference ontologies can help align and provide common grounding for mapping different information domains into one another [12]. Vetted, gold-standard reference structures provide a fixity of coherent touchpoints for orienting different concepts and domains (and, we believe, data) to one another.

In the data integration context, master data models (and management, or MDM) attempt to provide common reference terms and objects to aid the integration effort. Like other areas in conventional data integration, very few examples of MDM tools based on semantic technologies exist.

This use of reference structures and the importance of knowledge bases to help solve hard computational tasks suggests there may be a general principle at work. If ontologies can help orient domain concepts, why can't they also be used to orient instance data and their attributes? In fact, must these structures always be ontologies? Are not other common reference structures such as taxonomies, vocabularies, reference entity sets, or other typologies potentially useful to data integration?

By standing back in this manner and asking these broader questions we can see a host of structures like reference concepts, reference attributes, reference places, reference identifiers, and the like, playing the roles of providing common groundings for integration and interoperation. Through the AI experience, we can also see that subsequent use of these reference structures -- be they full knowledge bases or more limited structures like taxonomies or typologies -- can further improve information extraction and organization. The virtuous circle of knowledge structures improving AI algorithms, which can then further improve the knowledge structures, has been a real *Aha!* moment for the artificial intelligence community. We should see rapid iterations of this virtuous circle in the months to come.

These perspectives can help lead to purposeful designs and approaches for attacking such next-generation problems as data interoperability. The semantic Web can not solve this alone because additional AI capabilities need to be brought to bear. Conventional data integration approaches that lack semantic Big Structure groundings -- let alone the use of AI techniques -- have years of history of high cost and disappointing results. No conventional enterprise knowledge management problem appears sheltered from this whirlwind of knowledge-backed AI.

At Structured Dynamics, Fred Giasson and I have been discussing "Big Structure" for some time. However, it was only in researching this article that I came across the first public use of this phrase in the context of AI and big data. In May, Dr. Jiawei Han, a leading researcher in data mining, gave a lecture at Yahoo! Labs entitled, *Big Data Needs Big Structure*. In it, he defines "Big Structure as a type information network." The correlation with ontologies and knowledge structures is obvious.

An Emerging Development Agenda

The intellectual foundations already exist to move aggressively on a focused development agenda to improve the infrastructure of data interoperability. This emerging agenda needs to look to new refererence structures, better tooling, the use of functional languages and practices, and user interfaces and workflows that improve the mappings that are the heart of interoperability.

Big Structure, such as UMBEL for referencing what data is about, is the present exemplar for going forward. Excellent reference and domain ontologies for common domains already exist. Mapping predicates have been developed for these purposes. Though creation of the maps is still laborious, tooling improvements (see below) should speed up that process as well.

What is next needed are reference structures to help guide attributes mappings, data value mappings, and transformations into usable common attribute quantities and types. I will discuss in a later post our more detailed thoughts of what a reference gold-standard attribute ontology should look like. This new Big Structure should also be helpful in guiding conversion, transformation and "lifting" utilities that may be used to bring attribute values from heterogeneous sources into a common basis. As mappings are completed, these too can become standard references as the bootstrapping continues.

Mappings for data integration across the scales, scope and growth of data volumes on the Web and within enterprises can no longer be handled manually. Semi-automated tooling must be developed and refined that operates over large volumes with acceptable performance. Constant efforts to reduce the data volumes requiring manual curation are essential; AI approaches should be incorporated into the virtuous iterations to reduce these efforts. Meanwhile, attentiveness to productive user interfaces and efficient workflows are also essential to improve throughput.

Further, by working off of standards-based Big Structures, this tooling can be made more-or-less generic, with ready application to different domains and different data. Because this tooling will often work in enterprises behind firewalls, standard enterprise capabilities (security, access, preservation, availability) should also be added to this infrastructure.

These Big Structures and tools should themselves be created and maintained via functional programming languages and DSLs specifically geared to the circumstances at hand. We want languages suited to RDF and AI purposes with superior performance across large mapped datasets and unstructured text. But we also want languages that are easier to use and maintain by knowledge workers themselves. Partitioning strategies may also need to be employed to ensure acceptable real-time feedback to users responsible for data integration mappings.

A New Adaptive Infrastructure for Data Interoperability

Structured Dynamics' review exercise, now documented in this two-part series, affirms the semantic Web needs to become re-embedded in artificial intelligence, backed by knowledge bases, which are themselves creatures of the semantic Web. Coupling artificial intelligence with knowledge bases will do much to improve the most labor-intensive stumbling blocks in the data integration workflow: mappings and transformations. Through a purposeful approach of developing reference structures for attributes and data

values, we will begin to see marked improvements in the efficiency and lower costs of data integration. In turn, what is learned by using these approaches for mastering MDM will teach the semantic Web much.

An approach using semantic technologies and artificial intelligence tools will begin to solve the data integration puzzle. By leveraging background knowledge, we will begin to extend into data interoperability. Purposeful attention to tooling and workflows geared to improve the mapping speed and efficiency by users will enable us to increase the stable of reference structures -- that is, Big Structure -- available for the next integration challenges. As this roster of Big Structures increases, they can be shared, allowing more generic issues of data integration to be overcome, freeing domains and enterprises to target what is unique.

Achieving this vision will not occur overnight. But, based on a decade of semantic Web experience and the insights being gained from today's knowledge-based AI advances, the way forward looks pretty clear. We are entering a fundamental new era of knowledge-based computation. We welcome challenging case examples that will help us move this vision forward.

NOTE: This Part II concludes the series with Part I, A Decade in the Trenches of the Semantic Web

---

[1] Using semantic ontologies can and has worked well for many domains and applications, such as the biomedical OBO ontologies, IBM's Watson, Google's Knowledge Graph, and hundreds in more specific domains. Combined with concept reference structures like UMBEL, both building blocks and exemplars exist for how to interoperate across what different domains are about.

[2] For examples of issues, see M. K. Bergman, 2009. When Linked Data Rules Fail, *AI3:::Adaptive Information* blog, November 16, 2009.

[3] Some of these options are overviewed by M. K. Bergman, 2010. The Nature of Connectedness on the Web, *AI3:::Adaptive Information* blog, November 22, 2010.

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

[5] See QUDT - Quantities, Units, Dimensions and Data Types Ontologies, Retrieved July 22, 2014.

[6] The *object* may also refer to another class or instance, in which case the relation property takes the form of an ObjectProperty and the "value" is the URI referring to that object.

[7] See, for example, M. K. Bergman, 2009. Making Linked Data Reasonable Using Description Logics, Part 2, *AI3:::Adaptive Information* blog, February 15, 2009.

[8] See David Karger, 2013. Keynote at the European Semantic Web Conference Part 1: The State of End User Information Management, June 5, 2013.

[9] Info-Tech Research Group, 2011. Vendor Landscape Plus: Data Integration Tools, 72 pp.

[10] Gartner estimates that the data integration tool market was slightly over $2 billion at the end of 2012, an increase of 7.4% from 2011. This market is seeing an above-average growth rate of the overall enterprise software market, as data integration continues to be considered a strategic priority by organizations. See Eric Thoo, Ted Friedman, Mark

A. Beyer, 2013. Magic Quadrant for Data Integration Tools, research Report G00248961 from Gartner, Inc., 17 July 2013; see: http://www.gartner.com/technology/reprints.do?id=1-1HBEFSF&ct=130717&st=sb

[11] See M. K. Bergman, 2014. Spring Dawns on Artificial Intelligence, *AI3:::Adaptive Information* blog, June 2, 2014.

[12] See M. K. Bergman, 2011. In Search of 'Gold Standards' for the Semantic Web, *AI3:::Adaptive Information* blog, February 28, 2011.

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