Pragmatic Semantics for the Web of Data

Stefan Schlobach and Wouter Beek
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081HV Amsterdam
The Netherlands

Abstract

The success of the Web of Data (WOD) is based on the thorough understanding of, and agreement upon, the semantics of data and ontologies. But the Web of Data as a whole is complex, and inherently messy, contextualised, opinionated, in short: it is a market-place of ideas, rather than a database. Existing paradigms are inappropriate for dealing with this new type of knowledge structures.

The urgency of dealing with the non-standard characteristics of the Web of Data has been recognised, and separate initiatives try to tackle its individual manifestations, e.g. inconsistencies, contexts, vagueness, provenance, etc. Tomorrow's Web of Data requires novel semantics with efficient (generic) implementations to ensure semantic clarity, reuse and interoperability.

We recently introduced pragmatic semantics as a new semantic paradigm integrating elements from market theory and classical semantics into a framework of optimisation over truth-orderings, each representing a particular world-view. We propose nature-based algorithms to implement those semantics. We recently started a new research project, called PraSem, with the goal of investigating Pragmatic Semantics both from a theoretical and practical perspective.

Introduction

The Web of Data (WOD) connects data in a similar way as the WWW connects documents. Atomic data-units called resources are connected via typed links with arbitrary resources anywhere on the Web, and together these RDF triples form a gigantic graph of linked data. The meaning of the types can be fixed using standardised schema and ontology languages such as RDFS and OWL. The semantics of these languages are based on logical paradigms that were designed for small and hand-made knowledge bases, and come with a classical model-theory assigning truth to formulae, and entailment based on this truth. In a highly complex, dynamic, context-dependent, opinionated, contradictory and multi-dimensional semantic network as the WOD, these Semantics are insufficient, as they are one-dimensional, often prone to logical fallacies, and usually intractable.

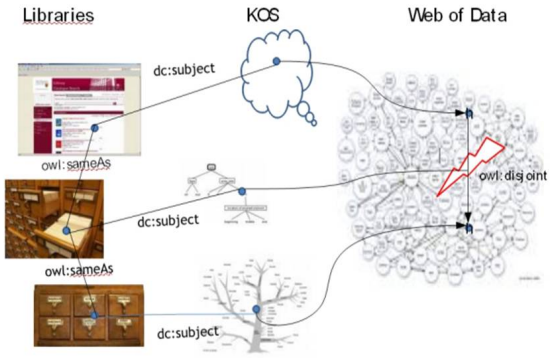

Figure 1: Scenario 1: heterogeneous publishing

Two simple scenarios will illustrate some of the high-level problems.

Scenario 1 is about heterogeneous publishing of data. Many libraries describe their books with controlled vocabularies. Linking collections and those vocabularies to the Linked-Open Data cloud, a collection of hundreds of interconnected data-graphs on the WOD, has huge benefits for libraries as search becomes more powerful, and metadata of documents is automatically enriched. Suppose a library in China annotates a book about Amsterdam with a concept ch:SmallTown. The Dutch National Library, on the other hand, annotates the same book with subject nl:BigCity. What happens now when the two libraries add their vocabularies and data to the WOD? What should be the desired answer to a query for big cities? Linking the libraries' vocabularies to the Linked-Open Data cloud will lead to conflicts and hamper access to the document in question rather than support it.

Scenario 2 is about opinionated interpretation of data, and is taken from Scientometrics, the Science of measuring and predicting Science. Scientometric researcher often use the Web as a proxy for studying science itself. Scientists leave online traces while doing research and a lot of this data is structured and part of the Web of Data. In some way,

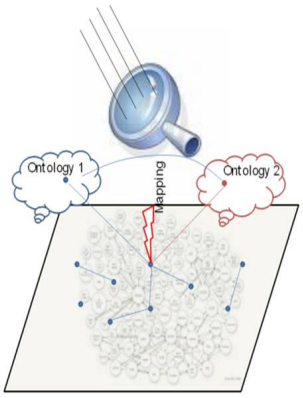

Figure 2: Scenario 2: opinionated data interpretation

the WOD becomes a magnifying glass to measure activity in Science. The problem, however, is that multiple views are omnipresent: research blogs are biased, there are networks of publications of different impact levels, social networks that overshadow reliable analysis, which all comes on top of the usual technical problems of instance- unification, homonymy and synonymy. Modelling this highly complex Science Web with standard ontology languages is impossible as long as standard semantics are enforced.

On the Web in general, and the Web of Data in particular, almost every bit of information is context-dependent, biased towards a particular viewpoint, opinionated, dated, uncertain or vague. The WOD is a market-place of ideas, not a database, and has to be dealt with accordingly. As making the representational languages more complex is not an option, we have to adapt the formal semantics of existing formalisms to the new requirements.

The need for a novel Semantic Paradigm

The goal of our new research project, PraSem will be to introduce and assess the potential of novel semantics that can help overcome the weakness of traditional semantics when dealing with a messy, multi-dimensional, contextualised and complex knowledge structure such as the Web of Data.

Consider a prototypical example: a Dutch dataset describes European cities, among them Amsterdam, which is a capital and does not require a visa for travel. In good practice the resources are linked to existing sources, e.g. DBpedia, by an owl:sameAs predicate. Similar data is published in China (using the namespace ch:), but now for European cities a visa is required. Both pieces of information are locally correct and the linking follows the correct principles. Still, considering the two classes ch:VisumNeeded and nl:VisumFreeCity to be disjoint, classical semantics collapse, and even useful derivable information, such as the fact that Amsterdam is a city with an airport, as it is a capital city, is lost. To address this issue, we need to deal with truth at different contexts.

Recent approaches to extend current methods, e.g., with

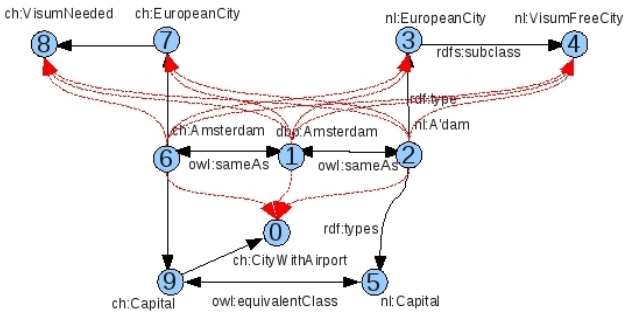

Figure 3: An example ontology

quantitative information about vagueness and uncertainty, or towards multi-dimensionality are highly useful for specific applications. But they necessarily fall short of representing the full rich of the Web of Data. They also fail on a second, critical, requirement for our new semantics: as current ontology languages are already perceived as being too complex for practical use, the burden cannot be put on the modeller. We need to adapt the semantics, not the languages.

More concretely, we consider ontologies to be sets of RDF(S)/OWL triples. Without loss of generality an interpretation consists of a domain and interpretation function, assigning individuals to objects, concepts and classes to subsets of this domain, and properties and roles to binary relations, and which extended over the operators of the underlying representation language. Models are interpretations satisfying all the axioms. Axioms or triples are then classically entailed by the ontology if they are satisfied in all its models. Unfortunately, even in our simple example things go wrong as there cannot be a model where the instance dbp:Amsterdam is both visa-free and a city where a visa is required. By definition, everything is entailed: the semantics becomes useless.

Pragmatic Semantics

Pragmatic semantics integrate different world-views instead of defining meaning with respect to a single one. The idea is to make as much information in the data explicit, and turn it into first-class semantics citizens. First, it allows integrating classic model-theoretic notions of truth with explicit knowledge about the structure of the knowledge base. But also semantic meta-data, such as popularity, scarcity, abnormality, etc., and even background knowledge from other sources, can be integrated.

Most of this additional knowledge induces some kind of ordering on the formulas, which we will call truth orderings. A simple example of such a truth ordering is the one induced by the size of the minimal subontology classically entailing a formula. Other examples are orderings derived as the ratio of sub-models (models for parts of an ontology) in which a formula is satisfied versus the total number of sub-models, or the ratio between sub-ontologies of O in which a formula holds holds versus the number of all sub-ontologies are interesting candidates. Those are orderings based on subset of

the ontology, a well-known class often used when dealing with inconsistent ontologies (Huang, van Harmelen, and ten Teije 2005).

Another relevant class are orderings induced by the graph properties of the ontology, in case that the underlying data-model is graph-based. A shortest path ordering can be determined as the inverse of the longest shortest distance between all nodes in the ontology (diameter of the induced sub-graphs). Such a notion is a proxy for confidence of derivation. Other graph-based measures, e.g. based on random-walk distance or edge-weights, induce orderings that are clustering-aware; with sub-ontologies entailing a formula have more cohesion than others. Finally, taking node properties such as PageRank into account, orderings can be used as proxies for popularity.

While those two classes of orderings make structural properties of the ontology explicit and use them to implicitly contextualise meaning, others are based on external information outside the ontology itself. Examples for this are the Google count and similarity (Cilibrasi and Vitanyi 2007) based on frequency of labels of resources on the WWW.

The different orderings cover different aspects of the "true" semantics of the Web of Data. To combine those aspects pragmatic entailment is defined through multi-objective optimisation. A pragmatic closure C for an ontology O and orderings f1 to fn is then a set of formulas that is Pareto-optimal (Pareto 1890) w.r.t. the optimisation problem $max\{f1(C),,fn(C)\}$.

Interoperability is then achieved by enriching an ontology with meta-information about semantic orderings, as well as agreement on the weighting of orderings. As there are possibly several pragmatic closures (different solutions on the Pareto-front) also agreement on the weighting of features is required. We will refer to the entailment induced by a given set of orderings as an instantiation of the family of pragmatic semantics.

Calculi for Pragmatic Semantics

Another way of looking at it is that the Web of Data is a Complex System (Gueret et al. 2011), with interlinked information at different scales of abstraction. A well-argued claim in the Complex Systems literature suggests that it is impossible to construct logical systems that capture the full meaning of a true Complex System (Bar-Yam 2005).

Results from studying the Web of Data as a Complex System show that considering different scales and levels of interactions make it impossible to engineer a web-scale reasoner (whatever the semantics considered), as traditional, decomposition-based approaches, are doomed with bandwidth limitations between the coordinating components (i.e. the datasets). Traditional semantics deal with this problem by an intrinsic reduction of the complexity: only one world-view, one perspective is considered at the time, the Web of Data is seen as a database. With pragmatic semantics, this advantage gets lost, and the computational price has to be paid, which applies that classical top-down reasoning becomes impossible.

It is often claimed that such systems have to evolve according to biological evolution rules (Bar-Yam 2004), and web-scale semantics and reasoning should emerge from controlled interactions between autonomous components. In (Dentler, Schlobach, and Guéret 2009) we introduced such a calculus based on swarm intelligence where instead of indexing all triples and joining the results, swarms of lightweight agents (so-called boids) autonomously traverse the graph, each representing a reasoning rule, which might be (partially) instantiated. Whenever the conditions of a rule match the node a boid is on, it locally adds the new derived triple. This provides an index-free alternative for reasoning over large distributed dynamic networks of RDF(S) graphs. It calculates the pragmatic closure under the condition of maximising popularity of nodes (as random walks of boids simulate PageRank calculation) and minimizing the length of sub-ontologies, two particular truth orderings. Not all of the conceivable calculi for pragmatic semantics have to be inspired by Computational Intelligence approaches, but PraSem will focus on this family of algorithms.

Related Work

The existing Semantic Web knowledge representation formalisms have been originally developed for describing crisp and static knowledge about a domain of application, and as such are essentially incapable of dealing with various contextual aspects of knowledge on the Web of Data, nor with a number of phenomena such as uncertainty, vagueness, ambiguity, which are a commonplace. On recognizing these limitations, much of the research effort of the Semantic Web community has been devoted to finding adequate ways of handling the newly identified tasks, resulting in a rich and heterogeneous body of work, e.g. on:

- Reasoning with multiple ontologies, such as Integration and modularization, contextualisation and temporalisation. For the first e-Connections (Kutz et al. 2004; Grau, Parsia, and Sirin. 2006) and Distributed Description Logics (Borgida and Serafini 2003) are typical examples. The extensive work on contextualization mostly provides extensions to DL and OWL-DL languages for representing contexts and context-dependent knowledge explicitly (Benslimane et al. 2006; Bouquet et al. 2003; Goczyla, Waloszek, and Waloszek 2007) as does (Huang and Stuckenschmidt 2005) with temporalization.
- Reasoning with imperfect knowledge: non-standard extensions to the DL languages for representing uncertainty and vagueness, e.g. (Qi, Pan, and Ji 2007; Lukasiewicz and Straccia 2008).
- Reasoning with multiple RDF graphs: extensions to the RDF framework based on formal aspects of multi-dimensionality, such as named graphs (Carroll et al. 2005), networked graphs (Schenk and Staab 2008) and multidimensional graphs (Gergatsoulis and Lilis 2005). Also related: Ontology matching for semantic interoperability (Doan and Halevy 2005; Kalfoglou and Schorlemmer 2005; Euzenat and Shvaiko 2007) for some overviews, and more specifically for context dependency of matching (Fenza, Loia, and Senatore. 2009; Albertoni and Martino. 2008; Duchateau, Bellahsene, and Roche. 2007)).

Common to those approaches is their attempt to represent the complexity of the information explicitly, and as such put the burden on the shoulders of the user. All this work however indicates that devising semantics for this more complex and contextualised information is extremely hard, even when various aspects of the complexity are treated independently.

The most relevant attempt to introduce a new semantic paradigm has been Emergent Semantics (Cudre-Mauroux 2009), defining semantics as the result of collective processes and interactions between nodes in a network - a collective agreement. Although this formalism can capture some of the emerging structure, the price is that meaning and truth are defined as results of processes or calculi, and the well-understood declarative, model-theoretic semantics of traditional formalisms are lost. PraSem is orthogonal to emergent semantics: it can be seen as an attempt to explicitly capture as much semantic information as possible. For this, the semantic properties need to be captured in the truth functions, and the complexity be dealt within the optimisation process. For developing truth functions we will study the extensive literature on Complex Systems and graphs (Newman, Barabsi, and Watts 2006; Wang and Groth 2010).

Very significant work on collective intelligence has proven to be effective for tasks such as network routing (Dorigo and Caro 1998) and data clustering (Deneubourg et al. 1990), and using nature-inspired methods has become standard for optimization problems (Deb 2001; Coello Coello, Lamont, and Van Veldhuizen 2007). They have also been investigate in the context of the Web of Data recently, such as for storage and querying (Mühleisen, Walther, and Tolksdorf 2011; Oren, Guéret, and Schlobach 2008), but to the best of our knowledge PraSem is novel in its attempt to use such methods as calculi for explicit semantic systems.

References

Albertoni, R., and Martino., M. D. 2008. Asymmetric and context-dependent semantic similarity among ontology instances. *Journal on Data Semantyics*.

Bar-Yam, Y. 2004. About engineering complex systems: Multiscale analysis and evolutionary engineering. In *Engineering Self-Organising Systems*, 16–31.

Bar-Yam. 2005. *Making Things Work: Solving Complex Problems in a Complex World*. Knowledge Press.

Benslimane, D.; Arara, A.; Falquet, G.; Maamar, Z.; Thiran, P.; and Gargouri, F. 2006. Contextual ontologies: Motivations, challenges, and solutions. In *Advances in Information Systems Conference*.

Borgida, A., and Serafini, L. 2003. Distributed description logics: Assimilating information from peer sources. *Journal of Data Semantics*.

Bouquet, P.; Giunchiglia, F.; van Harmelen, F.; Serafini, L.; and HeinerStuckenschmidt. 2003. C-owl: Contextualizing ontologies. In *ISWC*.

Carroll, J. J.; Bizer, C.; Hayes, P.; and Stickler, P. 2005. Named graphs, provenance and trust. In *WWW '05*, 613–622.

Cilibrasi, R., and Vitanyi, P. 2007. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering* 19(3):370–383.

Coello Coello, C. A.; Lamont, G. B.; and Van Veldhuizen, D. A. 2007. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer.

Cudre-Mauroux, P. 2009. Emergent semantics. *Encyclopedia of Database Systems* 982–985.

Deb, K. 2001. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons.

Deneubourg, J. L.; Goss, S.; Franks, N.; Sendova-Franks, A.; Detrain, C.; and Chretien, L. 1990. The dynamics of collective sorting robot-like ants and ant-like robots. In *In Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats.*, 356–363. MIT Press.

Dentler, K.; Schlobach, S.; and Guéret, C. 2009. Semantic web reasoning by swarm intelligence. In *5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009)*.

Doan, A., and Halevy, A. Y. 2005. Semantic integration research in the database community: A brief survey. *AI Magazine* 26(1).

Dorigo, M., and Caro, G. D. 1998. Antnet: Distributed stigmergetic control for communications networks. *Artificial Intelligence Research* 9:317–365.

Duchateau, F.; Bellahsene, Z.; and Roche., M. 2007. Context-based measure for discovering approximate semantic matching between schema elements. In *Proceedings of RCIS*.

Euzenat, J., and Shvaiko, P. 2007. *Ontology Matching*. Springer.

Fenza, G.; Loia, V.; and Senatore., S. 2009. Local semantic context analysis for automatic ontology matching. In *Proceedings of IFSA-EUSFLAT*.

Gergatsoulis, M., and Lilis, P. 2005. Multidimensional rdf. In Meersman, R., and Tari, Z., eds., *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, 1188–1205.

Goczyla, K.; Waloszek, W.; and Waloszek, A. 2007. Contextualization of a DL knowledge base. In *DL*.

Grau, B. C.; Parsia, B.; and Sirin., E. 2006. Combining owl ontologies using e-connections. *Journal of Web Semantics* 4(1):40–59.

Gueret, C.; Wang, S.; Groth, P. T.; and Schlobach, S. 2011. Multi-scale analysis of the web of data: a challenge to the complex system's community. *Advances of Complex Systems* 14(4):587–609.

Huang, Z., and Stuckenschmidt, H. 2005. Reasoning with multi-version ontologies: A temporal logic approach. In *ISWC*, 398–412.

Huang, Z.; van Harmelen, F.; and ten Teije, A. 2005. Reasoning with inconsistent ontologies. In *IJCAI*, 454–459.

Kalfoglou, Y., and Schorlemmer, W. M. 2005. Ontology mapping: The state of the art. In *Semantic Interoperability and Integration*, volume 04391 of *Dagstuhl Seminar Proceedings*.

Kutz, O.; Lutz, C.; Wolter, F.; and Zakharyaschev., M. 2004. E-connections of abstract description systems. *Artificial Intelligence* 156(1):1–73.

Lukasiewicz, T., and Straccia, U. 2008. Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics* 6(4):291–308.

Mühleisen, H.; Walther, T.; and Tolksdorf, R. 2011. Multilevel indexing in a distributed self-organized storage system. In *IEEE Congress on Evolutionary Computation*. 989–994.

Newman, M.; Barabsi, A.-L.; and Watts, D., eds. 2006. *The Structure and Dynamics of Networks*. Princeton University Press.

Oren, E.; Guéret, C.; and Schlobach, S. 2008. Anytime query answering in rdf through evolutionary algorithms. In *International Semantic Web Conference*. Springer. 98–113.

Pareto, V. 1890. The new theories of economics. *Journal of Political Economy* 5:485–502.

Qi, G.; Pan, J. Z.; and Ji, Q. 2007. Extending description logics with uncertainty reasoning in possibilistic logic. In *ECSQARU '07*, 828–839. Springer-Verlag.

Schenk, S., and Staab, S. 2008. Networked graphs: a declarative mechanism for sparql rules, sparql views and rdf data integration on the web. In *WWW '08*, 585–594.

Wang, S., and Groth, P. 2010. Measuring the dynamic bidirectional influence between content and social networks. In *ISWC*.