Michelle Cheatham
Data Semantics Lab, Wright State University

Tim Berners-Lee originally envisioned an Internet through which both humans and machines could share information, which he coined the semantic web [1]. Building on a set of standards including URIs, RDF, and OWL, the semantic web has been quietly growing for over a decade. There are now over 30 billion RDF triples (i.e. facts) in the linked open data cloud, covering everything from weather data, to traffic patterns, to background knowledge such as that found on Wikipedia. In addition, much more data is available as semi-structured documents such as HTML tables, Excel files, and PDF documents. Incorporating this wealth of data into decision-making processes would be invaluable. This is easier said than done, however. One issue is that the concept of the "data silo" is alive and well, even on the semantic web. While there are many data sets available, the number of links between them continues to grow at a very slow rate. What is needed is a set of tools and techniques to find and ingest relevant data sets into applications and research platforms. This would enable geoscience researchers to more effectively generate, model, and test hypotheses that involve data from a variety of subdomains.

There has been a lot of research into semantic data integration. It began in the context of database schema integration, but this is limited because such algorithms implicitly make a "closed world" assumption that is not always valid. For the last decade, researchers in semantic data integration have come together to compete against one another on data and schema alignment tasks in the Ontology Alignment Evaluation Initiative (OAEI). However, performance in the OAEI competitions has arguably plateaued [2]. Furthermore, current alignment algorithms are focused almost exclusively on finding simple 1-to-1 equivalence relationships, and much of the success on this can be attributed to basic string matching [4]. Even worse, the best-performing alignment systems are only capable of accurately finding these basic relationships between classes (i.e. concepts), not between the properties that act as the glue between those concepts [3]. A new approach is needed.

Enter Ontology Design Patterns (ODPs). The goal of an ontology design pattern is to represent the key classes and the relationships between them necessary to model a concept that recurs in many different domains, such as "Person" or "Project". A good ODP makes only the minimum ontological commitments necessary to describe the concept, i.e. it models everything involved in making the concept what it is, and nothing else. Developing an ODP is generally a collaborative process, in which domain experts provide expertise about the concept and data modelers ensure that the pattern will support the desired queries in a computationally efficient manner. An example can be found in [5].

How are ontology design patterns useful for semantic data integration? Part of the reason ontologies have failed to live up to the hype so far is because it is very difficult to align existing data sets to large monolithic ontologies that attempt to represent entire domains. The people who developed those data sets and ontologies had different backgrounds, different application goals, and generally different ways of seeing the world, which invariably results in logical contradictions during the alignment process. Because ODPs are limited in both scope and ontological commitments, these problems are greatly reduced. Additionally, the limited size of ODPs makes it more computationally feasible to find complex alignments between data sets and patterns. Indeed this approach has been a frequent topic of conversation at recent conferences, though it has yet to be explored. And, although this has also yet to be tried, we hypothesize that patterns can be adorned with synonyms and typical domain and range values in order to recognize occurrences of relevant data in semi-structured text.

An ODP-based approach to semantic data integration also has several side benefits. Design patterns are readily understandable to domain experts, so systems based upon such models can generate computationally meaningful output. Additionally, a knowledge base can use existing tools and techniques, such as logical reasoners, to highlight inconsistencies and missing information.

[1] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific American 284.5 (2001): 28-37.

[2] Cheatham, Michelle, and Pascal Hitzler. "Conference v2. 0: An uncertain version of the OAEI Conference benchmark." The Semantic Web–ISWC 2014. Springer International Publishing, 2014. 33-48.

[3] Cheatham, Michelle, and Pascal Hitzler. "The properties of property alignment." Proceedings OM-2014, The Ninth International Workshop on Ontology Matching, at the 13th International Semantic Web Conference, ISWC. 2014.

[4] Cheatham, Michelle, and Pascal Hitzler. "String similarity metrics for ontology alignment." The Semantic Web–ISWC 2013. Springer Berlin Heidelberg, 2013. 294-309.

[5] Vardeman, Charles, Adila Krisnadhi, Michelle Cheatham, et. al. “An Ontology Design Pattern for Material Transformation.” Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns, Riva del Garda, Italy, October 19, 2014.