Semantics and Ontologies For EarthCube

Gary Berg-Cross[1]*, Isabel Cruz[2], Mike Dean[3], Tim Finin[4], Mark Gahegan[5], Pascal Hitzler[6], Hook Hua[7], Krzysztof Janowicz[8], Naicong Li[9], Philip Murphy[9], Bryce Nordgren[10], Leo Obrst[11], Mark Schildhauer[12], Amit Sheth[6], Krishna Sinha[13], Anne Thessen[14], Nancy Wiegand[15], and Ilya Zaslavsky[16]

[1] SOCoP
[2] University of Illinois at Chicago
[3] Raytheon BBN Technologies
[4] University of Maryland, Baltimore County
[5] University of Auckland, New Zealand
[6] Wright State University
[7] NASA Jet Propulsion Laboratory
[8] University of California, Santa Barbara
[9] University of Redlands
[10] Rocky Mountain Research Station, USDA Forest Service
[11] MITRE
[12] NCEAS at University of California, Santa Barbara
[13] Virginia Tech
[14] Marine Biological Laboratory
[15] University of Wisconsin-Madison
[16] San Diego Supercomputer Center

Abstract. Semantic technologies and ontologies play an increasing role in scientific workflow systems and knowledge infrastructures. While ontologies are mostly used for the semantic annotation of metadata, semantic technologies enable searching metadata catalogs beyond simple keywords, with some early evidence of semantics used for data translation. However, the next generation of distributed and interdisciplinary knowledge infrastructures will require capabilities beyond simple subsumption reasoning over subclass relations. In this work, we report from the EarthCube Semantics Community by highlighting which role semantics and ontologies should play in the EarthCube knowledge infrastructure. We target the interested domain scientist and, thus, introduce the value proposition of semantic technologies in a non-technical language. Finally, we commit ourselves to some guiding principles for the successful implementation and application of semantic technologies and ontologies within EarthCube.

The semantic annotation of data and semantics-enabled search in metadata catalogs are part of many scientific workflow systems, e.g., Kepler [1]. In the past, semantic technologies have shown great potential in many biologically-focused cyberinfrastructures for both data annotation and semantic translation. There is some preliminary evidence to suggest that similar approaches would also add value in the geosciences, e.g., in the context of GEON. However, there appears to be some confusion about the role that

* Author names are listed in alphabetic order.

semantics can play within distributed next-generation knowledge infrastructures such as NSF's EarthCube[1]. Indeed, current Semantic technologies require knowledge of formal logic that is unfamiliar to most Earth scientists. There is, however, a simple way to understand how semantics can contribute greatly to the interoperability [2] of data, models, and services within EarthCube: simply put, by linking scientific observations and other data to terms drawn from ontologies or other forms of vocabularies, one can gain insights from how those terms are linked to other definitions in the ontology. This all happens *behind the scenes*. For example, if a scientist has collected observations of salinity measurements from the sea surface at location *X*, she can automatically link the data to terms like: chemical concentrations, oceanographic measurements, measurements (e.g., sea surface temperature) from 0m depth, and correlated measurements from locations situated near to *X* – all become accessible through the potential relationships revealed through ontologies. Thus, scientists searching for those general terms are more likely to find and potentially reuse the data. This capability will be invaluable to any scientist doing integrative or synthetic research that benefits from finding complementary data that others (e.g. potential collaborators) might have collected [3]. Even more, in an interdisciplinary setting the same terms may have different meanings and data may be collected and published following different measurement procedures and scientific workflows. Ontologies help to make such hidden heterogeneities explicit and, thus, support scientists in understanding whether a certain dataset fits their models [4]. Finally, to a certain degree, ontologies can also automatically translate data to make them interoperable and also reveal differences in the used classification systems [5].

If EarthCube promotes common vocabularies for annotating and describing data using terms drawn from ontologies, the value added will far exceed what can be expected from annotation using simple metadata, or worse, annotation using completely uncontrolled and not structured vocabularies. All the formal semantic processing and reasoning will be automatically accomplished behind the scenes for the scientists, in the same way that a Web browser nicely renders a page for a human to read. As a research community, we need to learn to be flexible, to develop techniques for *hardening* ontologies from looser semantics, to infer connections to more formal semantics, more generally to start with what is available whilst encouraging the development of more formal semantics where it is practical to do so. Google, Apple, the New York Times and Best Buy all use ontologies to support their content management systems or for other purposes related to sharing and managing of data. Thus, we believe that EarthCube should use semantic technologies as well. A key benefit of adopting Semantic technologies is that a vast number of repositories, ontologies, methods, standards, and tools that support scientists in publishing, sharing, and discovering data, is already available.

Semantic technologies provide new capabilities for formally and logically describing scientific facts and processes that may be as transformative as the introduction of the relational model was for organizing and accessing data over the past three decades. While a number of exciting semantic technology developments are underway, perhaps the area with greatest immediate applicability to EarthCube is the Semantic Web. The Semantic Web is a research field that studies how to foster the publishing, sharing, dis-

[1] See http://www.nsf.gov/geo/earthcube/ and the community page at http://earthcube.ning.com/ .

covery, reuse, and integration of data and services in heterogeneous, cross-domain, and large-scale infrastructures. It consists of two major components.

(i) Ontologies and knowledge representation languages that restrict the interpretation of domain vocabulary towards their intended meaning and, thus, allow us to conceptually specify scientific workflows, procedures, models, and data, i.e., the body of knowledge in a given domain, in a way that reduces the likelihood of misunderstanding and fosters retrieval and reuse [6].

(ii) As these ontologies are formal theories, they enable reasoning services on top of them. These reasoning services assist at different stages. They ensure that the developed ontologies are consistent. They help to make implicit knowledge explicit, discover incompatibilities and, thus, prevent users from combining data, models and tools that were developed with different underlying assumptions in mind. They allow querying across different sources and the semi-automatic alignment of different ontologies to foster the reuse and integration of data, models, and services. And finally, they support the design of smart user interfaces that go beyond simple keyword search and improve accuracy in search, cross-domain discovery, and other tasks which require data and information integration.

Linked Data is the data infrastructure of the Semantic Web [7]. It has rapidly grown over the last years and has found substantial uptake in industry and academia, since it significantly lowers the barrier for publishing, sharing, and reuse of data. Linked Data is an easily adoptable and ready-to-use paradigm that enables data integration and interoperation by opening up data silos. Combining Semantic Web technologies and Linked Data with ontologies also enables the discovery of new knowledge and the testing of scientific hypotheses. Consequently, the Semantic Web allows for vertical and horizontal integration, which is of central importance for EarthCube in order to realize the required interoperability of data, models, and tools while preserving the heterogeneity that drives the motor of interdisciplinary science.

However, the use of semantic technologies and ontologies in itself does not automatically guarantee interoperability or better access to data if not supported by a clear roadmap and guiding principles. The following list reflects a minimal set of principles that should guide the community for the next years. For EarthCube to be successful and transformative, we propose the following lines of action:

1. Be driven by concrete use cases and needs of the members of the EarthCube community. Collect, at the outset, a set of use cases from each EarthCube group, and conduct a substantial study of interconnected use cases which expose requirements related to data, models, and tools interoperability. These requirements need to be thoroughly analyzed as to the requirements they impose on the EarthCube data, ontology, and semantics infrastructure.
2. The choice of methods and the degree of knowledge formalization, e.g., lightweight versus heavyweight approaches, should be chosen based on use cases and application needs. This reduces the entry barrier for domain scientists to contribute data and ensures that a semantics-driven infrastructure is available for use in early stages of EarthCube.
3. Foster semantic interoperability without restricting the semantic heterogeneity introduced by the diverse community representing EarthCube. Provide methods that

enable users to flexibly load and combine different ontologies instead of hardwiring data to particular ontologies and, thus, hinder their flexible reusability.

4. Allow for bottom-up and top-down approaches to semantics to ensure a vertical integration from the observations-based data level up to the theory-driven formalization of key domain facts.
5. Involve domain experts in ontology engineering and enable them to become active participants by providing building blocks, strategies, documentations, and workshops on how to publish, retrieve, and integrate data, models, and workflows.
6. Apply semantics and ontologies to capture the body of knowledge in various Earth science domains for the purpose of organizing and accessing data, models and tools, learning about them, and extracting information from legacy data.
7. Exploit the power of classical and non-classical reasoning services to develop user interfaces, dialog systems and service chains that assist domain scientists at different stages ranging from discovering data and integrity constraint checking to the generation of new knowledge and hypothesis testing.

A detailed, more technical argumentation why these points need to be realized and how the heterogeneity of the geosciences requires new directions of research beyond schema standardization, can be found in the report of the Semantics and Ontology Technical Committee Report [8].

Acknowledgments

Krzysztof Janowicz acknowledges support by the National Science Foundation EarthCube Charette June 2012 travel grant. Pascal Hitzler acknowledges support by the National Science Foundation under award 1017225 "III: Small: TROn – Tractable Reasoning with Ontologies," and under a travel grant to the June 2012 EarthCube Charette. Anne Thessen acknowledges funding by NSF Data Net Program 0830976 and the National Science Foundation EarthCube Charette June 2012 travel grant.(c) 2012, by the respective authors, and The MITRE Corporation (for Leo Obrst). All Rights Reserved.

References

1. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience **18**(10) (2006) 1039–1065
2. Obrst, L.: Ontologies for semantically interoperable systems. In: Proceedings of the 12th international conference on Information and knowledge management. CIKM ’03, ACM (2003) 366–369
3. Jones, M., Schildhauer, M., Reichman, O., Bowers, S.: The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics **37**(1) (2006) 519–544
4. Janowicz, K., Hitzler, P.: The Digital Earth as knowledge engine. Semantic Web Journal **3**(3) (2012) 213–221

5. Gahegan, M., Smart, W., Masoud-Ansari, S., Whitehead, B.: A semantic web map mediation service: interactive redesign and sharing of map legends. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies. SSO '11, ACM (2011) 1–8
6. Kuhn, W.: Semantic Engineering. In Navratil, G., ed.: Research Trends in Geographic Information Science. Lecture Notes in Geoinformation and Cartography, Springer (2009) 63–76
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22
8. Hitzler, P., Janowicz, K., Berg-Cross, G., Obrst, L., Sheth, A., Finin, T., Cruz, I.: Semantic Aspects of EarthCube. Technical report, Semantics and Ontology Technical Committee. Available online at: http://knoesis.wright.edu/faculty/pascal/pub/EC-SO-TC-Report-V1.0.pdf (2012)