Semantic e-Science

Xiaogang Ma • Peter Fox • Thomas Narock • Brian Wilson

Received: 29 January 2015 / Accepted: 5 February 2015 / Published online: 15 February 2015
© Springer-Verlag Berlin Heidelberg 2015

Introduction

Jim Gray described e-Science as where “IT meets scientists.” (Hey et al. 2009) Science has now fully entered this new mode of operation, which combines science, informatics, computer science, cyberinfrastructure and information technology. It has been six years since the special issue Geoscience Knowledge Representation in Cyberinfrastructure (Brodaric et al. 2009) appeared in the journal Computers & Geosciences. In the ensuing years e-Science has changed how science disciplines conduct both individual and collaborative work. It is time to once again review the state of e-Science research. A special session was held at the American Geophysical Union (AGU) 2013 Fall Meeting.

Communicated by: H. A. Babaie

Published in the Special Issue of *Semantic e-Science* with Guest Editors Dr. Xiaogang Ma, Dr. Peter Fox, Dr. Thomas Narock and Dr. Brian Wilson

---

X. Ma (✉) • P. Fox
Tetherless World Constellation, Rensselaer Polytechnic Institute,
110 8th Street, Troy, NY 12180, USA
e-mail: max7@rpi.edu

P. Fox
e-mail: pfox@cs.rpi.edu

T. Narock
Department of Information Technology and Management Science,
Marymount University, 1000 Glebe Road, Arlington, VA 22201,
USA
e-mail: tnarock@marymount.edu

B. Wilson
Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Drive, Pasadena, CA 91109, USA
e-mail: bdwilson@jpl.nasa.gov

This special session, titled Semantically Enabling Annotation, Discovery, Access, and Integration of Scientific Data, hosted 25 presentations on current e-Science projects. We initiated this special issue by sending invitations to authors in the 2009 Computers & Geoscience special issue as well as the 2013 AGU presenters. Submission to this special issue was also open to everyone, and we were happy to have received manuscripts from authors across the world. This finalized special issue consists of 11 papers, which cover various subjects in Earth and environmental sciences, and demonstrate state-of-the-art technologies in knowledge representation, data interoperability, vocabulary and data services, and data processing.

Semantic e-Science

As e-Science flourishes and the barriers to data are being lowered, other more challenging questions are emerging, such as, “How do I use this data that I did not generate?” or “How do I use this data type, which I have never seen, together with the data I use every day?” or “What should I do if I really need data from another discipline but I cannot understand its terms?” Along with the growth in the volume, complexity, and heterogeneity of data resources, scientists increasingly need new capabilities that rely on semantic approaches (e.g., in the form of ontologies and vocabularies—machine encodings of terms, concepts, and relations among them) to help understand the meaning of data.

The field of semantic e-Science fosters the growth and development of data-intensive scientific applications based on semantic methodologies and technologies, as well as related knowledge-based approaches (Fox and Hendler 2009). In recent years, semantic methodologies and technologies have been gaining momentum in e-Science areas such as solar-

terrestrial physics, geology, ecology, oceanography, meteorology, and life sciences, to name a few. The developers of e-Science infrastructures are increasingly in need of semantic-based methodologies, tools, and middleware. This infrastructure will in turn facilitate scientific knowledge modeling, logic-based hypothesis checking, semantic data integration, application composition, integrated knowledge discovery and data analysis for different scientific domains, and building systems for use by scientists, students, and, increasingly, non-experts.

Modeling and encoding

Ontologies and vocabularies are the primary components of semantic technologies. There are various languages and schemas used in the Semantic Web, such as the Resource Description Framework (RDF), the Web Ontology Language (OWL) and the Simple Knowledge Organization System (SKOS). However, the work developing Earth science ontologies and vocabularies is not a re-design of the wheel. Cox and Richard show that geologic time conceptual models in Unified Modeling Language (UML) can be used to generate both eXtensible Markup Language (XML) schemas and OWL ontologies, which in turn can be used to build SKOS vocabularies. A good conceptual model lasts a long time and can be encoded in various languages. Abel et al. express a similar point of view on the modeling and encoding issues. But they extend the discussion, by using examples in the petroleum system, to how to use ontologies to reconcile the conceptual models developed by different geologists.

Community of practice to promote data interoperability

To address data heterogeneities at the semantic level, more communities of practice are needed. The aim of developing ontologies and vocabularies is to have a common language for both computers and humans. An ideal mechanism of creating, curating and using ontologies and vocabularies is from the community, by the community and for the community. Diviacco et al. discuss the methodology of a boundary object used in the European Commission Eurofleets project. The boundary object is resulted from discussions among the ‘divergent’ communities. It embeds core and shared conceptual entities and is used across those communities to address heterogeneity. Duerr et al. have stakeholders participate in a use case driven iterative approach for developing a family of ontologies for sea ice. The output ontologies also leverage existing class models for characterizing sea ice to improve the compatibility. Wright et al. develop a thesaurus for describing environmental chemistry datasets, and they plan to open the thesaurus to the user community for further extension, which

includes both concepts in the thesaurus itself and links to external vocabularies.

Rethinking the data life cycle

Conventionally, we consider a data life cycle begins with data collection, continues with processing, archiving, distribution, discovery, analysis, and then repurposing. From repurposing, the life cycle may go back to the processing step restarting the cycle. Recent studies (Ma et al. 2014) show that before data collection there is another step called concept, which includes works on the conceptual model, ontology and terminology of datasets. Knowledge of the ontology and terminology has many implications for the ensuing steps in the data life cycle. Ji et al. present a vocabulary service as a semantic support to the Environmental Data Store. The service hosts other commonly used vocabularies in the field of environmental studies. Wright et al. develop a method to use a developed thesaurus for tagging datasets. The interoperability among the tagged datasets is thus promoted, which also facilitates the data discovery and repurposing in later stages. Li et al. develop a semantic matrix in a search tool for intelligent discovery of polar datasets. The tool can identify hidden semantic associations between terminologies in the datasets metadata and store them in the matrix. The matrix is maintained persistently and improves search performance.

Semantics for relationship recognition and inference

A key feature differentiates semantic technologies from conventional technologies is their capability for relationship recognition and semantic inference. Fiorini et al. propose a framework for representing similarity and part-whole relations. They exemplify the framework by demonstrating an algorithm that supports cognitive processes in stratigraphic interpretation in well log datasets. Many Earth science datasets are recorded in natural language, and natural language processing is also used in the semantic web. The Agenames developed by Huber and Klump offers an online geological text parser. In a given text, Agenames can identify stratigraphic terms and use these terms to infer and assign a geologic age estimate. Zheng et al. develop the Information Entropy based Weighted Similarity Model for computing similarity among concepts and suggesting possible links. Besides texts, semantic technologies can also be used in image processing. He et al. develop ontologies for complex geospatial features and enrich them with fuzzy sets of spatial relations. They deploy the ontologies in a geoprocessing service environment and demonstrate on-demand uncertainty-aware detection of complex geospatial features from remote sensing images.

Concluding remarks

The papers in this special issue cover various subjects from stratigraphy, marine science, polar science, to geography etc. and present different aspects of the semantic technologies. Although we assemble them into the above four clusters, our short introduction just reveals a part of the characteristics of each paper. We encourage you to read the full text of any paper that is of interest to you for more details. Semantic technologies are increasingly discussed and used in Earth and environmental sciences, and we hope this special issue can provide a vision of progress in recent years, and offer the methodology and technology that will be beneficial to your work.

Acknowledgments We gratefully thank the authors for their excellent manuscripts and reviewers for constructive comments and suggestions. We also want to thank the managing editor Ms. Jenny Diego and the

Editor-in-Chief Dr. Hassan Babaie for coordinating the communication and helping us keep the timeline.

References

- Brodaric B, Fox P, McGuinness DL (2009) Geoscience knowledge representation in cyberinfrastructure. *Comput Geosci* 35(4):697–699
- Fox P, Hendler J (2009) Semantic eScience: encoding meaning in next-generation digitally enhanced science. In: Hey T, Tansley S, Tolle K (eds) *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, Redmond, pp 147–152
- Hey T, Tansley S, Tolle K (2009) Jim Gray on eScience: a transformed scientific method. In: Hey T, Tansley S, Tolle K (eds) *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, Redmond, pp xvii–xxxi
- Ma X, Fox P, Rozell E, West P, Zednik S (2014) Ontology dynamics in a data life cycle: challenges and recommendations from a Geoscience Perspective. *J Earth Sci* 25(2):407–412