



# *Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science*

PETER FOX

JAMES HENDLER

Rensselaer Polytechnic  
Institute

SCIENCE IS BECOMING INCREASINGLY DEPENDENT ON DATA, yet traditional data technologies were not designed for the scale and heterogeneity of data in the modern world. Projects such as the Large Hadron Collider (LHC) and the Australian Square Kilometre Array Pathfinder (ASKAP) will generate petabytes of data that must be analyzed by hundreds of scientists working in multiple countries and speaking many different languages. The digital or electronic facilitation of science, or eScience [1], is now essential and becoming widespread.

Clearly, data-intensive science, one component of eScience, must move beyond data warehouses and closed systems, striving instead to allow access to data to those outside the main project teams, allow for greater integration of sources, and provide interfaces to those who are expert scientists but not experts in data administration and computation. As eScience flourishes and the barriers to free and open access 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, 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?” This list of questions is large and growing as data and information product use increases and as more of science comes to rely on specialized devices.



An important insight into dealing with heterogeneous data is that if you know what the data “means,” it will be easier to use. As the volume, complexity, and heterogeneity of data resources grow, scientists increasingly need new capabilities that rely on new “semantic” approaches (e.g., in the form of ontologies—machine encodings of terms, concepts, and relations among them). Semantic technologies are gaining momentum in eScience areas such as solar-terrestrial physics (see Figure 1), ecology,<sup>1</sup> ocean and marine sciences,<sup>2</sup> healthcare, and life sciences,<sup>3</sup> to name but a few. The developers of eScience infrastructures are increasingly in need of semantic-based methodologies, tools, and middleware. They can in turn facilitate scientific knowledge modeling, logic-based hypothesis checking, semantic data integration, application composition, and integrated knowledge discovery and data analysis for different scientific domains and systems noted above, for use by scientists, students, and, increasingly, non-experts.

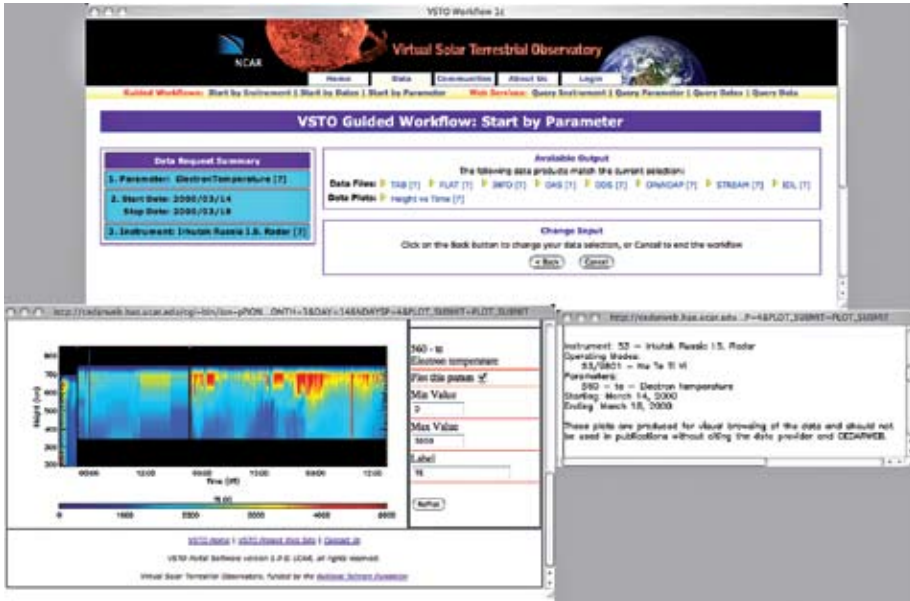
The influence of the artificial intelligence community and the increasing amount of data available on the Web (which has led many scientists to use the Web as their primary “computer”) have led semantic Web researchers to focus both on formal aspects of semantic representation languages and on general-purpose semantic application development. Languages are being standardized, and communities are in turn using those languages to build and use ontologies—specifications of concepts and terms and the relations between them (in the formal, machine-readable sense). All of the capabilities currently needed by eScience—including data integration, fusion, and mining; workflow development, orchestration, and execution; capture of provenance, lineage, and data quality; validation, verification, and trust of data authenticity; and fitness for purpose—need semantic representation and mediation if eScience is to become fully data-intensive.

The need for more semantics in eScience also arises in part from the increasingly distributed and interdisciplinary challenges of modern research. For example, the availability of high spatial-resolution remote sensing data (such as imagery) from satellites for ecosystem science is simultaneously changing the nature of research in other scientific fields, such as environmental science. Yet ground-truthing with *in situ* data creates an immediate data-integration challenge. Questions that arise for researchers who use such data include, “How can ‘point’ data be reconciled with various satellite data—e.g., swath or gridded—products?” “How is the spatial

<sup>1</sup> E.g., the Science Environment for Ecological Knowledge (SEEK) and [2].

<sup>2</sup> E.g., the Marine Metadata Interoperability (MMI) project.

<sup>3</sup> E.g., the Semantic Web Health Care and Life Sciences (HCLS) Interest Group and [3].



**FIGURE 1.** The Virtual Solar-Terrestrial Observatory (VSTO) provides data integration between physical parameters measured by different instruments. VSTO also mediates independent coordinate information to select appropriate plotting types using a semantic eScience approach without the user having to know the underlying representations and structure of the data [4, 5].

registration performed?” “Do these data represent the ‘same’ thing, at the same vertical (as well as geographic) position or at the same time, and does that matter?” Another scientist, such as a biologist, might need to access the same data from a very different perspective, to ask questions such as, “I found this particular species in an unexpected location. What are the geophysical parameters—temperature, humidity, and so on—for this area, and how has it changed over the last weeks, months, years?” Answers to such questions reside in both the metadata and the data itself. Perhaps more important is the fact that data and information products are increasingly being made available via Web services, so the semantic binding (i.e., the meaning) we seek must shift from being at the data level to being at the Internet/Web service level.

Semantics adds not only well-defined and machine-encoded definitions of vo-



cabularies, concepts, and terms, but it also explains the interrelationships among them (and especially, on the Web, among different vocabularies residing in different documents or repositories) in declarative (stated) and conditional (e.g., rule-based or logic) forms. One of the present challenges around semantic eScience is balancing expressivity (of the semantic representation) with the complexity of defining terms used by scientific experts and implementing the resulting systems. This balance is application dependent, which means there is no one-approach-fits-all solution. In turn, this implies that a peer relationship is required between physical scientists and computer scientists, and between software engineers and data managers and data providers.

The last few years have seen significant development in Web-based (i.e., XML) markup languages, including stabilization and standardization. Retrospective data and their accompanying catalogs are now provided as Web services, and real-time and near-real-time data are becoming standardized as sensor Web services are emerging. This means that diverse datasets are now widely available. Clearinghouses for such service registries, including the Earth Observing System Clearinghouse (ECHO) and the Global Earth Observation System of Systems (GEOSS) for Earth science, are becoming populated, and these complement comprehensive inventory catalogs such as NASA's Global Change Master Directory (GCMD). However, these registries remain largely limited to syntax-only representations of the services and underlying data. Intensive human effort—to match inputs, outputs, and preconditions as well as the meaning of methods for the services—is required to utilize them.

Project and community work to develop data models to improve lower-level interoperability is also increasing. These models expose domain vocabularies, which is helpful for immediate domains of interest but not necessarily for crosscutting areas such as Earth science data records and collections. As noted in reports from the international level to the agency level, data from new missions, together with data from existing agency sources, are increasingly being used synergistically with other observing and modeling sources. As these data sources are made available as services, the need for interoperability among differing vocabularies, services, and method representations remains, and the limitations of syntax-only (or lightweight semantics, such as coverage) become clear. Further, as demand for information products (representations of the data beyond pure science use) increases, the need for non-specialist access to information services based on science data is rapidly increasing. This need is not being met in most application areas.

Those involved in extant efforts (noted earlier, such as solar-terrestrial physics,



ecology, ocean and marine sciences, healthcare, and life sciences) have made the case for interoperability that moves away from reliance on agreements at the data-element, or syntactic, level toward a higher scientific, or semantic, level. Results from such research projects have demonstrated these types of data integration capabilities in interdisciplinary and cross-instrument measurement use. Now that syntax-only interoperability is no longer state-of-the-art, the next logical step is to use the semantics to begin to enable a similar level of semantic support at the data-as-a-service level.

Despite this increasing awareness of the importance of semantics to data-intensive eScience, participation from the scientific community to develop the particular requirements from specific science areas has been inadequate. Scientific researchers are growing ever more dependent on the Web for their data needs, but to date they have not yet created a coherent agenda for exploring the emerging trends being enabled by semantic technologies and for interacting with Semantic Web researchers. To help create such an agenda, we need to develop a multi-disciplinary field of *semantic eScience* that fosters the growth and development of data-intensive scientific applications based on semantic methodologies and technologies, as well as related knowledge-based approaches. To this end, we issue a four-point call to action:

- Researchers in science must work with colleagues in computer science and informatics to develop field-specific requirements and to implement and evaluate the languages, tools, and applications being developed for semantic eScience.
- Scientific and professional societies must provide the settings in which the needed rich interplay between science requirements and informatics capabilities can be realized, and they must acknowledge the importance of this work in career advancement via citation-like metrics.
- Funding agencies must increasingly target the building of communities of practice, with emphasis on the types of interdisciplinary teams of researchers and practitioners that are needed to advance and sustain semantic eScience efforts.
- All parties—scientists, societies, and funders—must play a role in creating governance around controlled vocabularies, taxonomies, and ontologies that can be used in scientific applications to ensure the currency and evolution of knowledge encoded in semantics.



Although early efforts are under way in all four areas, much more must be done. The very nature of dealing with the increasing complexity of modern science demands it.

#### REFERENCES

- [1] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-Science," *Science*, vol. 308, no. 5723, May 2005, pp. 817–821, doi: 10.1126/science.1110410.
- [2] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An Ontology for Describing and Synthesizing Ecological Observation Data," *Ecol. Inf.*, vol. 2, no. 3, pp. 279–296, 2007, doi: 10.1016/j.ecoinf.2007.05.004.
- [3] E. Neumann, "A Life Science Semantic Web: Are We There Yet?" *Sci. STKE*, p. 22, 2005, doi: 10.1126/stke.2832005pe22.
- [4] P. Fox, D. McGuinness, L. Cinquini, P. West, J. Garcia, and J. Benedict, "Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience," *Comput. Geosci.*, vol. 35, no. 4, pp. 724–738, 2009, doi: 10.1.1.141.1827.
- [5] D. McGuinness, P. Fox, L. Cinquini, P. West, J. Garcia, J. L. Benedict, and D. Middleton, "The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research," *AI Mag.*, vol. 29, no. 1, pp. 65–76, 2007, doi: 10.1145/1317353.1317355.