A Deployed Semantically-Enabled Interdisciplinary Virtual Observatory

Deborah McGuinness[1,2], Peter Fox[3], Luca Cinquini[4], Patrick West[3],
Jose Garcia[3], James L. Benedict [2] and Don Middleton[4]

[1] Knowledge Systems, Artificial Intelligence Laboratory, Stanford University, 345 Serra Mall, Stanford, CA 94305,
{dlm@cs.stanford.edu}
[2] McGuinness Associates, 20 Peter Coutts Circle, Stanford, CA 94305, {dlm,
jbenedict}@mcguinnessassociates.com
[3] High Altitude Observatory, Earth Sun Systems Lab, National Center for Atmospheric Research, PO Box 3000,
Boulder, CO 80307, {pfox,pwest,jgarcia}@ucar.edu
[4] Scientific Computing Division, Computing and Information Systems Lab, National Center for Atmospheric Research, PO Box 3000,
Boulder, CO 80307, {luca,don}@ucar.edu

Abstract

We have used semantic technologies to design, implement, and deploy an interdisciplinary virtual observatory. The Virtual Solar-Terrestrial Observatory is a production data framework providing access to observational datasets. It is in use by a community of scientists, students, and data providers interested in the middle and upper Earth's atmosphere, and the Sun. The data sets span upper atmospheric terrestrial physics to solar physics. The observatory allows virtual access to a highly distributed and heterogeneous set of data that appears as if all resources are organized, stored and accessible from a local machine. The system has been operational since the summer of 2006 and has shown registered data access by over 75% of the active community (last count over 600 of the estimated 800 person active research community). This demonstration will highlight how semantic technologies are being used to support data integration and more efficient data access in a multi-disciplinary setting. A full paper on this work is being published in the IAAI 07 'deployed' paper track.

1. Introduction

Scientific data is being generated, collected, and maintained in digital form in high volumes by many individual research groups. The need for access to and interoperability between these repositories is growing both for research groups to access their own data collections but also for researchers to access and utilize other research groups' data repositories in a single discipline or, more interestingly, in a multi-discipline manner. The promise of the true virtual interconnected heterogeneous distributed international data repository is starting to be realized. But there exist many challenges including interoperability and integration between data collections. We are exploring ways of technologically enabling scientific virtual observatories - distributed resources that may contain vast amounts of scientific observational data, theoretical models, and analysis programs and results from a broad range of disciplines. Our goal is to make these repositories appear as if they are one integrated local resource, while realizing that the information is collected by many research groups, using multiple instruments with varying settings in many experiments with different goals, and captured in a wide range of formats. Also, as interdisciplinary research grows, the range of the science areas is broadening and thus it is nearly impossible for users to have depth in all of the areas. We used artificial intelligence technologies, in particular semantic technologies, to create declarative, machine operational encodings of the semantics of the data to facilitate interoperability and semantic data integration. We then semantically enabled web services to find, manipulate, and present scientific data.

Our setting is interdisciplinary virtual observatories – data collections that cover multiple disciplines. Although the (current) typical user has a high level of training and education, almost no user has expert-level training in all of the disciplines covered in the collection. There are many challenges to providing an integrated retrieval interface including vocabulary differences across disciplines; varying customs for experimental settings of instruments, inconsistent or incomplete meta data, etc. These challenges become more significant when users want to go beyond search and manipulate and use the resulting data. These challenges expand significantly when we broaden our target user population to include high school and junior high teachers and the general public. Our primary approach to the interoperability problem is to encode formal specifications of terms and their interrelationships in an ontology. Our applications then use these machine operational descriptions of terms to inform their processing. These semantically-enabled applications can then be used directly or can be embedded in applications that provide customized interfaces appropriate for specialized communities. One primary value of the knowledge representation and reasoning is for supporting users (whether the users are programs or humans) with interfaces that find, access and use data in a more effective, robust and reliable way.

2. Demonstration

In our demonstration, we will show how a typical user interacts with the system and contrast it with how previous data query and access was performed. We will identify processing portions that are enabled by the underlying semantic technologies. We will also highlight some benefits that users are reporting from using our services.

Typical users have some idea of the kind of scientific data they want to retrieve but they may not know all of the data sets that contain the data (nor all the instruments used to obtain the data nor the observatories containing those instruments). In order to help users obtain relevant data in a more efficient manner, we are providing services that use background knowledge to formulate specific queries to the appropriate data collections to retrieve the data.

The general form of the use case is "retrieve data (from appropriate collections) subject to (stated and implicit) constraints and plot in a manner appropriate for the data". In the demonstration, we will show instantiations of types of that use case and highlight where the ontologies, reasoning, and web services are helping.

The initial motivating use case scenarios are provided below in a templated form and then in an instantiated form:
Template 1: Plot the values of parameter X as taken by instrument description or instance Y subject to constraint Z during the period W in style S.
Example 1: Plot the Neutral Temperature (Parameter) taken by the Millstone Hill Fabry-Perot interferometer (Instrument) looking in the vertical direction from January 2000 to August 2000 as a time series.
Template 2: Find and retrieve image data of the type for images of content Y during times described by Z.
Example 2: Find and retrieve quick look and science data for solar corona images during a recent observation period.
Template 3: Find data for parameter X constrained by Y during times described by Z.
Example 3: Find data representing the state of the neutral atmosphere anywhere above 100km and toward the Arctic circle (above 45N) at times of high geomagnetic activity.
Template 4: Expose semantically-enabled, smart data query services using constraint X and concepts Y_1, Y_2, etc. via a web services interface allowing composite query formation in arbitrary workflow order.
Example 4: Provide query services for the Virtual Ionosphere-Thermospere-Mesosphere Observatory that retrieve data filtered constraints on Instrument, Date-Time, and Parameter in any order and with constraints included in any combination.

3. AI Technology Usage Highlights

The demonstration will highlight some benefits of semantic technologies. We can give specific examples of

1. Decreased input requirements from the user:
2. Syntactically correct query support:
3. Semantically meaningful query support:
4. Semantic integration:
5. Broader range of potential users:

We attribute the benefits to the underlying technology. Additionally, we can show how better semantic integration allows retrieval of relevant data that the user may not have known about. Further, we can show how those additional retrievals can change the way science research may be done in the future.

The main AI elements that support the semantic foundation for integration in our application include the OWL ontologies, a description logic reasoner (along with supporting tool infrastructure for ontology editing and validation), and our semantically-enabled web services.

Some of our project contributions will be identified in the demonstration. Some include reusable and extensible general observational science ontologies (covering items such as instruments, observatories, etc.). We have already tested these ontologies in other science settings not directly related to atmospheric and solar-influences research including volcano, plate tectonics, and local, regional and global climate change research. Potentially of more interest is using our demonstration and methodology detailed in the IAAI paper as an operational prototype of a semantically-enabled interdisciplinary virtual observatory, or possibly more broadly construed as an operational prototype and methodology that may be of use in any project requiring interdisciplinary scientific data integration search, retrieval, and analysis.

Acknowledgements

References

D. McGuinness, P. Fox, L. Cinquini, P. West, J. Garcia, J.. Benedict, and D. Middleton. The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In the proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07). Vancouver, British Columbia, Canada, July 22-26, 2007.