Semantic Data Integration and Registration: Application to heterogeneous atmosphere and volcanic data sources.

Deborah McGuinness
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th St, Troy NY 12180
(1) 518-276-4404
dlm@cs.rpi.edu

Robert Raskin
JPL/NASA, 4800 Oak Grove,
Pasadena, CA 91109
(1) 818-354-4228
Robert.Raskin@jpl.nasa.gov

Peter Fox
HAO/ESSL/NCAR
PO Box 3000
Boulder, CO 80307
(1) 303-497-1511
pfox@ucar.edu

A Rezgui
Virginia Tech
4044 Derring Hall (0420)
Blacksburg, VA 24061
(1) 540-231-6521
arezgui@vt.edu

Krishna Sinha
Virginia Tech
4044 Derring Hall (0420)
Blacksburg, VA 24061
(1) 540-231-5580
pitlab@vt.edu

Patrick West
HAO/ESSL/NCAR
PO Box 3000
Boulder, CO 80307
(1) 303.497.1560
pwest@ucar.edu

ABSTRACT
We present our work on semantically-enabled data and schema registration in the setting of a scientific data integration project: SESDI (Semantically-Enabled Scientific Data Integration), which aims initially to integrate heterogeneous volcanic and atmospheric chemical compound data in support of assessing the atmospheric effects of a volcanic eruption. We use semantic methods through out the project, however in this paper and demonstration, we will highlight issues related to our work on data and schema registration and integration. In this process, we will demonstrate how we are re-using previously developed ontologies and how those ontologies are being used to provide a "smart: data integration capability aimed at interdisciplinary scientific research.

Categories and Subject Descriptors
H.2.5 {**Heterogeneous Databases**}: {Data translation}, I.2.4 [**Knowledge Representation Formalisms and Methods**]: Relation systems, Representation languages, Representations (procedural and rule-based), Semantic networks.

Keywords
Semantic data integration; ontologies – modular; data frameworks; semantic data registration; use case methodology.

1. INTRODUCTION
Our overall effort aims to enable the next generation of interdisciplinary and discipline-specific data and information systems. While we focus in this project on two specific areas (volcanology and atmospheric data), our goal is to provide an infrastructure that can be used in many natural science settings. The driving natural science focus in SESDI is to support research on relationships between volcanic activity and global climate (McGuinness et al. 2006, 2007, Fox et al. 2007a, b, Fox et al. 2008b, Sinha et al. 2006, 2007a, b). The driving computer science focus is on providing infrastructure for semantically-enabling data and schema integration in disparate domains.

This work is aimed at providing scientists with the option of describing what they are looking for in terms that are meaningful and natural to them, instead of in a syntax that is not. The goal is not simply to facilitate search and retrieval, but also to provide an underlying framework that contains information about the semantics of the scientific terms used. Our system is expected to be used by scientists who want to do processing on the results of the integrated data, thus the system must provide access to how integration is done and what definitions it is using. There are numerous online scientific data services, but the missing element in previous systems is declarative specification of term meanings and in supporting technologies leveraging ontologies and ontology-equipped tools, interfaces, and services. In the rest of this paper, we focus on our work on creating semantically aware interfaces between science components. This includes a focus on the registration of disciplinary data sets. To achieve this, we have developed a tool to aid data providers in registering data without explicitly knowing about the underlying ontologies.

2. SEMANTIC DATA INTEGRATION METHODOLOGY
Since we depend on machine processable specifications of the science terms that are used in the disciplines of interest, we are following a methodology reported in previous work (Benedict et al. 2007). We have identified specific ontology modules that need construction in the areas of volcanoes, plate tectonics, atmosphere, and climate drawing heavily on existing ontologies. In the present work, we utilize ontologies in the form of modules from SWEET (Semantic Web for Earth and Environmental Terminology), VSTO (Virtual Solar Terrestrial Observatory) and GEON (Geosciences Network). We scope our effort by focusing on the "Atmosphere-Volcano Use Case": " To determine the statistical signatures of both volcanic and solar forcings on the height of the tropopause" (or in lay person's terms, "to look for evidence of things that we can measure in the lower portion of the atmosphere that will impact human life"). We have been able to immediately leverage the VSTO framework (Fox et al. 2006, 2008, McGuinness et al. 2007) by replacing the solar-terrestrial-specific ontology and data sources with the appropriate volcano and atmospheric ontologies as mid-level ontologies tied together with our upper structure and by adding appropriate data/ catalog sources. We added Feature and Event classes, which are required to represent volcanoes, and eruptions, for example but do not

discuss this further in this report. This means the software built to support the VSTO application is re-used with new ontologies loaded and service classes added to communicate with the existing data sources.

Since we wanted to leverage community ontologies with wide adoption, we attempted to reuse SWEET 1.2 (http://sweet.jpl.nasa.gov) as much as possible. One interesting result of our effort is that we collaborated to help provide requirements for a much more modular (and much more reusable) version, now SWEET 2..0, which is to be published in August 2008. We helped obtain and facilitate broad community input for this modularization. In the demonstration, we will show how the modularized mid level community ontology has been leveraged in our system. In the demonstration, we will highlight the re-use of the previous semantic infrastructure components, modularization requirements for SWEET, and show how this was used in our data registration component.

3. DATA REGISTRATION

We base our data registration effort on the work from GEON (http://www.geongrid.org) and VSTO (http://www.vsto.org http://vsto.hao.ucar.edu). The data registration breaks sensibly into three levels (Baru et al. 2008):

Level 1. Discovery of data resources (e.g., gravity, geologic maps, etc) requires registration through use of high-level index terms. GEON has deployed extension of AGI Index terms that will be cross-indexed to others such as the Global Change Master Directory (GCMD) from NASA, the American Geophysical Union (AGU) as well as sub-disciplinary categorization

Level 2. Discovering Item level databases requires registration at data-type level ontologies (e.g. bulk rock geochemistry, gravity database)

Level 3. Item detail level registration (e.g., column in geochemical database that represents SO2 measurement). This level of registration is a requirement for semantic integration.

The demonstration will include an example data registration and show how some of those elements are required for the exemplar use case.

4. DISCUSSION AND CONCLUSION

We have developed a semantically-enabled scientific data integration demonstration that (a) leverages existing infrastructure and ontologies from the VSTO and Geon projects (b) motivated a modularization of a community-driven mid-level ontology and (c) serves as a pedagogical implemented system. Next steps include expanding the ontologies (around solar forcings) and expanded data registration prior to a community evaluation.

5. REFERENCES

[1] Baru, C., Fox, P. and Lin, K. 2007, The 1-2-3 of Data Registration, Earth Science Informatics, in preparation.

[2] Benedict, J. L.; McGuinness, D. L.; Fox, P. 2007, A Semantic Web-based Methodology for Building Conceptual Models of Scientific Information, Eos Trans. AGU, 88(52), Fall Meet. Suppl., Abstract IN53A-0950.

[3] Fox, P., D. L. McGuinness, L. Cinquini, P. West, J. Garcia, J. Benedict and D. Middleton, 2008, Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience, Computers and Geosciences, special issue on Geoscience Knowledge Representation for Cyberinfrastructure. in press.

[4] Fox, P, D L McGuinness, R Raskin, K Sinha, 2007a. A Volcano Erupts: Semantically Mediated Integration of Heterogeneous Volcanic and Atmospheric Data, CIMS'07, November 9, 2007, Lisboa, Portugal. Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience,

[5] Fox, P., Sinha, A.K. McGuinness, D., and Raskin, R. 2008b, Semantic Integration of Heterogeneous Volcanic and Atmospheric Data, Geoinformatics Conference 2008, San Diego, CA, U.S.Geological Survey Scientific Investigations Report 2007-5199, p. 10.

[6] Fox, P., Sinha, K., Raskin, R. McGuinness, D.L., Ammann, C., Venezky, D., Schwander, F. 2007b, Semantic Mediation and Integration of Volcanic and Atmospheric Data: In Search of Statistical Signatures, Eos Trans. AGU, 88(52), Fall Meet. Suppl., Abstract IN240A-05.

[7] McGuinness, D. L., 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. In the proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07). Vancouver, British Columbia, Canada, July 22-26, 2007.

[8] McGuinness, D. L., A. K. Sinha, P. Fox, R. Raskin, G. Heiken, C. Barnes, K. Wohletz, D. Venezky, K. Lin. Towards a Reference Volcano Ontology for Semantic Scientific Data Integration. American Geophysical Union Joint Assembly, Baltimore, Maryland, May 23-26, 2006.

[9] McGuinness, D. L. 2007, Semantic Integration of Heterogeneous Volcanic and Atmospheric Data, Geoinformatics Conference, San Diego, CA.

[10] Raskin, et al. 2007, Community Science Ontology Development Geoinformatics Conference, San Diego, CA. U.S.Geological Survey Scientific Investigations Report 2007-5199, p. 39.

[11] Sinha, A.K., Heiken, G., Barnes, C., Wohletz, K., Venezky, D., Fox, P., McGuinness, D.L, Raskin, R., and Lin,K, 2006, Towards an ontology for Volcanoes, U.S.Geological Survey Scientific Investigations Report 2006-5201, p.51

[12] Sinha, K., Raskin, R., McGuinness, D.L., Fox, P. 2007a Developing packages and integrating ontologies for Volcanoes, Plate Tectonics and Atmospheric Science Data Integration, Eos Trans. AGU, 88(52), Fall Meet. Suppl., Abstract IN53B-1204.

[13] Krishna Sinha, et al. 2007b, Towards a Reference Plate Tectonics and Volcano Ontology for Semantic Scientific Data Integration, Geoinformatics Conference, San Diego, CA. U.S.Geological Survey Scientific Investigations Report 2007-5199, p. 43.

ACKNOWLEDGMENTS

This work is funded by NASA/AIST+ACCESS. NCAR is operated by the UCAR with substantial sponsorship from the National Science Foundation.