A Volcano Erupts: Semantically Mediated Integration of Heterogeneous Volcanic and Atmospheric Data

Peter Fox
High Altitude Observatory,
ESSL/NCAR
PO Box 3000
Boulder CO 80307-3000
pfox@ucar.edu

Deborah McGuinness
Knowledge Systems & AI Lab,
Stanford University
353 Serra Mall
Stanford, CA 94305
McGuinness Associates
20 Peter Coutts Circle
Stanford, CA 94305

Robert Raskin
JPL/NASA
Wilson Blvd., Pasadena, CA

Krishna Sinha
Virginia Polytechnic Institute,
Department of Geology
Blacksburg, VA

ABSTRACT

We present a research effort into the application of semantic web methods and technologies to address the challenging problem of integrating heterogeneous volcanic and atmospheric data in support of assessing the atmospheric effects of a volcanic eruption.

This exemplary volcano eruption scenario highlights what is true for the vast majority of data intensive Earth system investigations which have limited ability to explore important and difficult problems. This is because they are forced to find and use data representing an event or phemenonon of interest through data collections at the data-element, or syntactic, level rather than at a higher scientific, or semantic, level. Even if relevant data is found in one collection, it may not be easy or possible to find similar, related data in another collection. In many cases, syntax-only interoperability IS the state-of-the-art and at best, there are some instances of hard-wired but simple semantic enhancements (e.g. a special purpose web service wrapper around the data). Scientists and non-scientists are forced to learn details of the data schema, other people's naming schemes and syntax decisions and details of differing web site interfaces. These constraints are limiting even when researchers are looking for information in their own discipline, but they present even greater challenges when researchers are looking for information spanning multiple disciplines, including some in which they are not extensively trained. The volcano eruption scenario exemplifies many of these challenges. In this paper we present research progress on how semantic enablement for scientific data integration is achieved. We present how ontologies implemented within existing distributed technology frameworks are providing essential, re-useable, and robust, support necessary for interdisciplinary scientific research activities.

CIMS'07, November 9, 2007, Lisboa, Portugal.

Categories and Subject Descriptors: H.2.5 Heterogeneous Databases: Data translation, I.2.4 Knowledge Representation Formalisms and Methods: Frames and scripts; Representation languages; Semantic networks

General Terms: Design

Keywords: informatics, knowledge representation, ontologies, semantic data integration, semantic mediation

1. INTRODUCTION

When a volcano erupts, there is sequence of events and impacts that is diverse and complex. The characteristics of an eruption; size, type and duration all influence the effect on the local, regional, and global environment. These effects range from diminished air quality, hazards for human health and ground and air transportation to effects on atmospheric composition and radiative blanketing. The contributions come from the smoke and ash, ejected gases, scattering and numerous other processes. The location of the volcano (latitude and longitude) as well as its tectonic setting on land or undersea also are factors. There are an increasing number of online repositories of scientific data information related to volcanoes, their present and past activity and both direct and proxy measurements of the nature of their impact.

While numerous sources of monitoring and retrospective data are available which represent measurements of the abovementioned quantities they are presently stored in heterogenous and highly distributed repositories. To realize the goal of integration of many of these diverse sources of data to address specific aspects of the volcano eruption scenarios we need to address many factors concerning access to and interoperability of the online scientific data.

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. These capabilities are 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. The missing elements in previous systems in enabling the higher-level semantic interconnections is the technology of ontologies, ontology-equipped tools, semantically aware interfaces between science components, and explanations of knowledge provenance. We present the initial results of a project entitled: Semantically-Enabled Science Data Integration (SESDI) [1] which uses semantic technologies to integrate data between these two discipline areas to assist in establishing causal connections as well as exploring as yet unknown relationships.

We use as starting points, many elements of semantic web methodologies and technologies which are based on our developments for the Virtual Solar-Terrestrial Observatory (VSTO) [2]. This work created a scalable environment for searching, integrating, and analyzing databases distributed over the Internet required a high level of semantic interoperability and has implemented a semantic data framework built on OWL-DL [3] ontologies, using the Pellet [4] reasoner within a Java-Tomcat servlet engine and made available via a Spring-based web portal and SOAP/WSDL [5] web services.

We also have significant experience with ontology packages and data registration from the Geosciences Network (GEON) [6]. Our present ontology develop involved some new developments as well as iterations and augmentations of the background domain ontology: Semantic Web for Earth and Environmental Terminology (SWEET) [7]. We leverage the precise formal definitions of the terms in supporting semantic search and interoperability.

2. USE CASES

We have developed several underlying use cases [8] for the volcano-atmosphere data integration and in this section we will present the first of these which addresses the signature of an eruption in the terrestrial lower atmosphere. This use case is: "determine the statistical signatures of both volcanic and solar forcings on the height of the tropopause".

This specific science template is motivated by the more general research direction of looking for indicators of the fall out of volcanic eruptions that may create changes in the atmosphere. The statistical signatures are such indicators, and the tropopause is at the edge of the atmosphere so it is a sensitive signature area to examine.

A schematic of the use case is shown in Fig. 1 which indicates some of the important terms, concepts, processes (and eventually underlying data) we need to represent.

3. METHODOLOGY

Our effort depends on machine processable specifications of the science terms that are used in the disciplines of interest. We are following a methodology that we believe is yielding candidate reference ontologies in our chosen domains. We have identified specific ontology modules that need construction in the areas of volcanoes, plate tectonics, atmosphere, and climate. We have begun construction of two of the modules along with the help of a set of selected

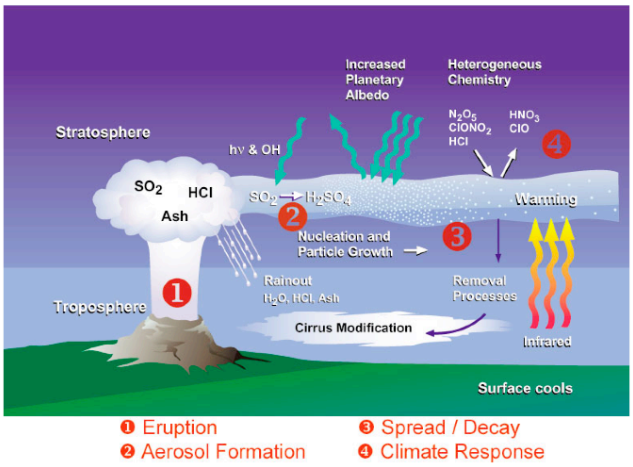

Figure 1: Schematic of the events and processes beginning with a volcanic eruption and leading to the climate/atmospheric response.

experts in the areas. Prior to a workshop, we identify a small set of subject matter experts. We also provide some background material for reading about ontology basics. Additionally, prior to our face to face meetings with experts, we identify foundational terms in the discipline and provide a simple starting point for organizing the basic terminology. While we do not want to influence the domain experts on their terminology, we find that we make more progress if we provide simple starting points using well agreed upon terminology. We then bring together a small group of the chosen domain experts and science ontology experts with a goal of generating an initial ontology containing the terms and phrases typically used by these experts. We use our task of researching the impact of volcanoes and global climate to focus the discussions to help determine scope and level of granularity.

We held a meeting with volcano experts and generated an initial ontology containing terms and phrases used to classify volcanoes, volcanic activities, and eruption phenomena. We use a graphical concept mapping tool [9] for capturing the terms and their relationships.

3.1 Volcanic and Plate Tectonic Semantics

Volcanoes can be classified by composition, tectonic setting, environmental setting, eruption type, activity, geologic setting, and landform. The ontology currently contains upper level terms in these areas and is being expanded according to the needs of the project and is being reviewed by additional domain experts. The initial focus is on gathering terms, putting them into a generalization hierarchy (using isa links in the diagram) and connecting the terms through properties (using has links in the diagram) as well as sameas and ispartof.

We held a second workshop to create a plate tectonics ontology. We identified domain experts and used the same science ontology experts as used in our volcano ontology meeting. We also focused in this meeting on gathering the primary class terms, e.g., plate boundary, lithosphere, etc., and putting them into a generalization hierarchy and identifying important properties relating the terms.

Both in preparation for and in follow-up from the domain workshops, we are reviewing relevant existing vocabularies and ontologies. We have reused terminology from SWEET, GEON ontologies, and the Virtual Solar-Terrestrial Observatory [10] instrument and observatory ontologies. We are also gathering some of the starting points for the atmosphere and climate ontologies from SWEET and related ontologies (the current version of the concept map for the atmosphere can be found at http://sesdi.hao.ucar.edu/cmaps/atmosphere_current.jpg).

We apply semantic web methodologies in pursuit of the SESDI objectives. These methods include the development and elaboration of use cases (user scenarios) from subject matter experts. In our project those experts are in volcanoes, plate tectonics and, atmospheric effects in response to forcings. We convene small workshop groups along these topic lines and start with use cases and elements of the existing vocabularies and/or ontologies where available and develop the knowledge representation using an interactive concept mapping tools (CMAP). During the first year of this project we held two of these workshops (volcanoes and plate tectonics) and another smaller workshop (to initiate the atmospheres work). The starting points going into these workshops and their nominal end-points (although not the end product) are a key indicator of our progress.

3.2 Atmospheric Semantics

Since our use case from above drives our need for the heterogeneous data integration we performed the same methodological approach to our knowledge representation of the atmospheric concepts and relations. The main difference is that we had as a starting point the SWEET. We had one of our domain literate team members extract key terms from SWEET and represent them into a concept map. We then engaged atmosphere science experts to refine and evolve the important concepts and relations in the concept map. Fig. 2 displays the results of one of the first iterations driven by the use case. In this figure, the concept elements added or altered are indicated with bolded borders.

In relation to climate effects, it is the far field forcing, i.e. non-intrinsic contributions, to altering the atmosphere and local, regional and global climate that is the first addition. At the top of Fig. 2, among the far field forcings are: TectonicSetting, VolcanicActivity, etc. In the center of the figure is the concept of the Tropopause (which is a atmospheric layer boundary) which has at least two properties; lower and upper boundary height, i.e. the signature of volcanic eruption forcing which is of interest in the use case.

Also of note is the indication that an atmospheric layer is part of climate and climate has primary substance(s), which include atomic consituents of the atmosphere (carbon, nitrogen, oxygen, and so on) as well as aerosols and contaminants - examples are SO_2, NO_x, ash, etc. The next phase of the project, which will be complete in the summer of 2007, will add the remaining properties, terms, relations and processes (an example of scattering as a physical process is in the aforementioned figure), based on the detail of the use case from the domain science expert. In the initial pass, we added 20% more terms to the ontology (which was already based on an extensive existing vocabulary).

We contributed new modules and expanded terms and concepts for the SWEET ontology in the case of the solid earth environment as well as additing relations to the atmospheric concepts; a key for our application..

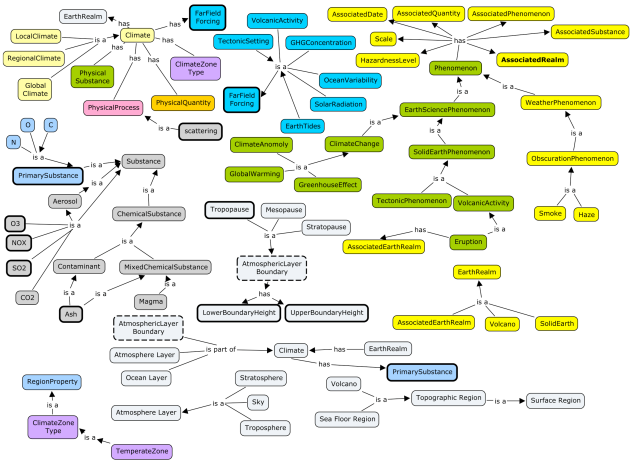

Figure 2: An excerpt of the atmosphere/climate concept map (after the workshop) with specific concepts that are motivated by the use case.

4. MEDIATING THE DATA INTEGRATION

4.1 Leveraging the Geosciences Network

The Geosciences Network project [6] has developed, among many cyberinfrastructure capabilities we will not discuss and do not utilize in this effort, a three step view of registering data [11] to enable discovery, access, and integration of heterogeneous data resources. Such a registration involves associating discovery, inventory and item/detail level metadata with the underlying datasets as a service that may be accessed from a data portal or invoked as a web service. The service generates registration metadata to facilitate inventorying, discovery, federation and integration of independent, heterogeneous data resources. Registering a data resource with a registration service does not require or imply that the data themselves are stored at a centralized location - though they could be. The data resources can be distributed.

The 3-step approach consists of:

1. Metadata Registration, where basic metadata about a resource is registered with the system. Metadata registration enables discovery of resources.

2. Schema registration, where schema elements of structured data resources are registered to an ontology, or a standard schema. Schema registration creates an inventory of resources with syntactic and structural descriptions of resources, and permits semi-automated integration of data across resources.

3. Data Item Registration, where individual data values in a data resources are registered to ontologies. With data item registration it is possible to provide very powerful data search engines and automated integration of data across heterogeneous resources

In addition, the DIA engine [12, 13, 14] is a Web services-based infrastructure for the Discovery, Integration, and Analysis (DIA) of geoscience data, tools, and services. DIA provides a collaborative environment for a data manager,

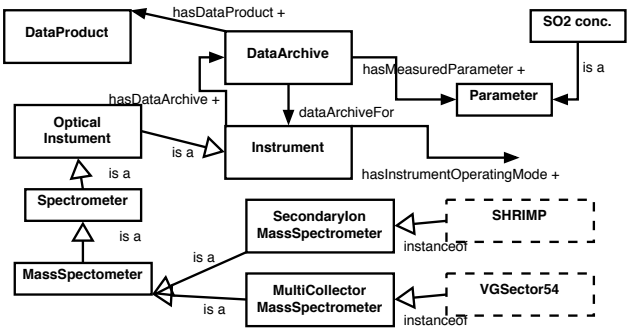

Figure 3: Schematic of VSTO Ontology 1.0 indicating a variety of classes: for data, service, service implementation and value restrictions. We also indicate a few properties/associations, inheritance and inference.

and/or scientists to share their resources (e.g., geochemical data, filtering services, etc.) by registering them through well-defined ontologies (based on a planetary materials ontology in OWL-DL). The ontology is used by different geoscientists (using Web services) to explore, extract, and integrate information from different heterogeneous data sets.

4.2 Leveraging the Virtual Solar-Terrestrial Observatory

The Virtual Solar-Terrestrial Observatory [2, 10] has developed a production semantic data framework in support of the solar, solar-terrestrial and space physics observational communitities. VSTO has a flexible and re-usable ontology which we have added to in this project; adding instrument sub-classes and instances and measured parameters relevant to our application areas of volcanoes and climate.

Fig. 3 is an excerpt of the VSTO instrument ontology with the addition of mass spectrometer which are instruments to measure parameters such as SO_2 concentration.

4.3 SWEET

In developing the atmospheric ontology, we drew upon the terms and concepts in the semantic web for earth and environmental terminology (SWEET) ontology. Our goal was to keep our ontology development separate until we believed it was stable and vetted at two different workshops which brought together domain scientists to discuss foundational earth and space science ontologies and related issues.

As noted earlier we used SWEET 1.0 to populate a concept map for the atmospheric ontology and updated it based on the use case. Based on our work in this project on the volcano ontology, plate tectonics ontology, as well from input from the GEON and VSTO projects. SWEET 1.1 was released in April of 2007 and will continue to evolve in response (version 1.2 is scheduled for August 2007) to modular/ package evolution of components of the ontologies.

The community portal for sharing the ontologies across these projects is the planetary ontologies web site[1], where ontologies can be downloaded, uploaded, and differences on different versions of ontologies can be performed. We encourage all community members to register, visit the website, download, compare and upload ontologies to contriute to the community dialog and process.

[1]http://www.planetont.org

4.4 Packaging the Ontology and Services

In conjunction with the post-workshop analysis which leads to the ontology development, we utilize a conceptual decomposition and modular approach (which is a best-practice in the semantic web methodology) to ontologies. Thus as we developed the classes and sub-classes in the volcano ontology, we associated them with one of the faceted or integrative ontologies indicated in the figure. Further, Fig. 4 is a schematic of such an approach for a volcano ontology and in particular how it leverages/ imports many other ontologies.

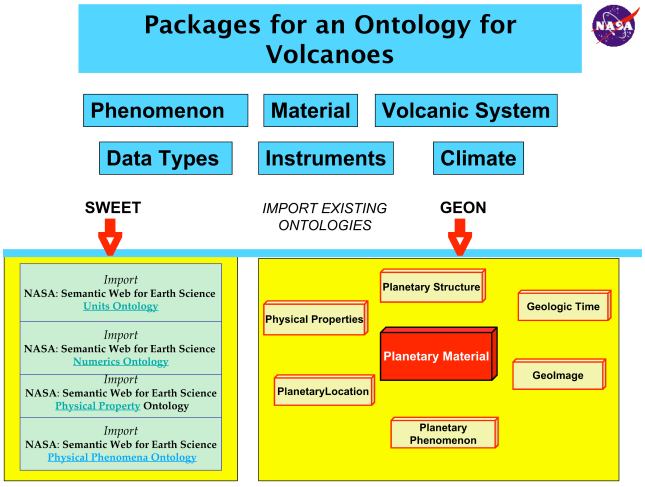

Figure 4: Packages approach to the development of a volcano ontology. This approach imports (re-uses) existing ontologies (Sinha, private communication).

4.5 Mapping to underlying data

Our candidate datasets for registration with the developed ontologies and for use in our data integration use case include the WOrld VOlcano DATabase (WOVODAT) [15] in collaboration with Yellowstone researchers and the USGS and the Nevada Test site database in collaboration with Los Alamos researchers. We are presently identifying the corresponding atmospheric and climate record databases. Our approach will be to establish initially a web portal based on one of the current best of-breed semantic data frameworks (e.g. VSTO, GEON) to prototype the access to the data from the volcano and atmosphere disciplines separately. Then we will provide web service access to both sources to that the statistical application we will utilize for the data integration can query for and retrieve data.

4.6 Benefits of a Semantic Web Approach

We have found numerous benefits from using the semantic web approach in our efforts to share and integrate information.

- We are finding that a broader range of users are able to use data services (largely because the interface does not require users to be experts in arcane instrument codes and access methods).

- We are also finding that the background ontologies are supporting more flexible web service support to the underlying data.
- Our representation of terms that are likely to be used in queries provides us with many more options for meaningful queries. Users may query the schema to find subclasses and instances in the ontology. This may be useful when they are forming their own detailed (potentially re-usable queries). Possibly one of the most interesting benefits is that users may retrieve information that they did not know was there. For example, our ontology supports a modeling of one instrument "acting as" another, thus if one is querying for information obtained by one instrument acting in another mode, the system will retrieve this as well. Many users would not know enough to specifically ask for this information if they needed to generate the explicit query themselves.
- We are finding that the upper level ontology classes, such as instrument and instrument properties are providing an excellent foundation for inheritance and expansion. One experience we had was convening the volcano ontology knowledge acquisition session and finding that we only needed to minimally expand our instrument ontology that was developed for solar and solar terrestrial physics. While of course we needed to add a few new instruments, we did not find the need for new properties nor new classes. The same experience was repeated in the plate tectonics ontology meeting and it was repeated again as we did the homework for the larger atmosphere ontology meeting.

A big advantage was gained from the inference that is possible to achieve the integration of heterogeneous data sources as follows:

Our web services wrappers provide a simple and extensible access method for querying and retrieving data from multiple sites. Users can obtain data according to any workflow order (as opposed to previous interfaces), they can retrieve at the class and/or instance level, but the most significant item is that they can query without being an expert in the individual schemas of the multiple data services.

5. DISCUSSION AND CONCLUSION

We have begun an effort that utilized ontologies to provide the capture term meanings in distinct but related science domains with a goal of facilitating research into relationships between the domains. We currently have starting points for reference ontologies in volcanoes and plate tectonics. We have also begun the homework on atmosphere and climate and will be holding workshops to generate reference ontologies with domain experts. We also will hold workshops to vet the ontologies among the multiple communities. Our findings so far are that our methodology for creating starting points for reference ontologies is working well in terms of gathering terms and relationships, and reaching agreement among the initial domain and science ontology experts. ‘

Based on the successful use of semantics in data integration for the VSTO and GEON projects, our next step for SESDI is to articulate a use case that drives the way and type of data integration needed to solve a specific scientific problem. Our candidate is to examine the statistical relation between the height of the tropopause and related forcings. This height is very sensitive to forcing and in a way that the fingerprint of volcanic and (for example) solar forcings are very distinct.

Acknowledgements

This work is supported by the SESDI project which is a semantic science data integration project sponsored by NASA Advancing Collaborative Connections for Earth-Sun System Science (ACCESS) and NASA Earth-Sun System Technology Office (ESTO) under award AIST-QRS-06-0016.

6. ADDITIONAL AUTHORS

Additional authors: Patrick West (HAO/ESSL/NCAR), Stephan Zednik (HAO/ESSL/NCAR), and James Benedict (McGuinness Associates).

7. REFERENCES

[1] Semantically-Enabled Science Data Integration - http://sesdi.hao.ucar.edu/, Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., Solomon, S. 2006, Semantically-Enabled Large-Scale Science Data Repositories. the 5th International Semantic Web Conference (ISWC06), LNCS, ed. Cruz et al., vol. 4273, pp. 792-805, Springer-Verlag, Berlin. Fox, P., McGuinness, D.L., Raskin, R. Sinha, A.K. 2006, Semantically-Enabled Scientific Data Integration. U.S. Geological Survey Scientific Investigations Report 2006-5201, (Geoinformatics 2006). 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 (Geoinformatics 2006). P. Fox, Deborah L. McGuinness, Rob Raskin, A. Krishna Sinha 2006, Semantically-enabled Science data Integration, Eos Trans. AGU 87(36), Jt. Assem. Suppl., Abstract IN42A-02. D.L. McGuinness, A.K. Sinha, P. Fox, R. Raskin, G. Heiken, C. Barnes, K. Wohletz, D. Venezky, K. Lin 2006, Towards a Reference Volcano Ontology for Semantic Scientific Data Integration, Eos Trans. AGU 87(36), Jt. Assem. Suppl., Abstract IN42A-03. Peter Fox, Deborah L. McGuinness, Rob Raskin, and A. Krishna Sinha 2006, The Technology Behind Data Integration with Semantics. Eos Trans. AGU 87(52), Fall Meet. Suppl., Abstract IN24A-05. Rob Raskin, Peter Fox, Deborah L. McGuinness, and A. Krishna Sinha 2006, Semantically-Enabled Science Data Integration: Current Progress. Eos Trans. AGU 87(52), Fall Meet. Suppl., Abstract IN43D-05. McGuinness, D. L., Fox, P., Sinha, A. K., and Raskin, R. 2007, Semantic Integration of Heterogeneous Volcanic and Atmospheric Data.: Proceedings of the Geoinformatics Conference, San Diego, CA., May 17-18, 2007, in press.

[2] Virtual Solar-Terrestrial Observatory - http://www.vsto.org, http://vsto.hao.ucar.edu, McGuinness, D. L., Fox, P., Cinquini, L., Darnell, J. A., West, P., Benedict, J. L., Garcia, J., and Middleton, D. 2006, Ontology-Enabled Virtual Observatories: Semantic Integration in Practice. Proc. of OWL Experiences and Directions 2006

(OWLED2006), CEUR Workshop Proceedings, vol. 216, online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-216/submission_14.pdf McGuinness, D. L., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J. L., and Middleton, D. 2007, 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.

[3] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. Available from http://www.w3.org/TR/owl-features/

[4] Pellet - http://www.mindswap.org/2003/pellet/

[5] Christensen, E., Curbera, F., Meredith, G., and Weerawarana, S. Web Services Description Language (WSDL) 1.1 - W3C Note 15 March 2001.

[6] Keller, G., Seber, D., Sinha, A.K. and Baru, C. 2005, The Geosciences Network (GEON): one step towards building cyberinfrastructure for the geosciences, European Geophysical Union, Geophysical Research Abstracts, Vol. 7, 05726, 2005 SRef-ID: 1607-7962/gra/EGU05-A-05726, http://www.cosis.net/abstracts/EGU05/05726/EGU05-J-05726.pdf, http://www.geongrid.org/

[7] Semantic Web for Earth and Environmental Terminologies - http://sweet.jpl.nasa.gov

[8] Cockburn, A., Writing Effective Use Cases, Addison-Wesley, Boston, MA, 2000.

[9] The Concept Mapping Ontology Editor - http://cmap.ihmc.us/coe

[10] Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J. and Middleton, D. 2007, Development of Solar-Terrestrial Ontologies for Semantic Scientific Data Frameworks, Computers and Geosciences, submitted.

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

[12] Malik, Z., Rezgui, A., and Sinha, A. K. 2007, Ontologic Integration of Geoscience Data on the Semantic Web, Proceedings of the Geoinformatics Conference, San Diego, CA., May 17-18, 2007, in press.

[13] Zaki M., A. Rezgui, A. K. Sinha, K. Lin, and A. Bouguettaya 2007, DIA: A Web Services-based Infrastructure for Semantic Integration in Geoinformatics, Proceedings of the IEEE ICWS 2007, Application Services and Industry Track, submitted.

[14] Rezgui, A., Malik, Z., and Sinha, A. K. 2007, DIA Engine: Semantic Discovery, Integration, and Analysis of Earth Science Data, Proceedings of the Geoinformatics Conference, San Diego, CA., May 17-18, 2007, in press.

[15] The World Volcano Database - http://www.wovo.org/WOVOdat