Integrating Inter-disciplinary Science Data with Semantic Mediation

Peter Fox (pfox@ucar.edu)
High Altitude Observatory, ESSL/NCAR, Boulder CO
Deborah McGuinness
Tetherless World Constellation, Rensselaer Polytechnic Institute
Troy, NY
McGuinness Associates, Stanford, CA
Robert Raskin
JPL/NASA, Pasadena, CA
Krishna Sinha
Virginia Polytechnic Institute, Department of Geology
Blacksburg, VA

Abstract

We present results of a research effort into the application of semantic web methods and technologies to address the challenging problem of integrating data from heterogeneous sources - in particular from volcanic and atmospheric chemistry data in support of assessing a particular science question: what are measureable atmospheric effects of of a volcanic eruption. The introduction of formal semantics in our methods and into the implemented technical infrastructure allows scientists to ask measurement based questions and retrieve relevant data rather than issuing instrument and data product specific oriented searches, often requiring very detailed and customized knowledge that is rarely replicable. One of the underlying principles is that scientists and non-scientists should not be forced to learn complex details of the data product naming and schema, other people's naming vocabularies, schemes and syntax decisions and myriad details of differing web site interfaces. The volcano eruption scenario exemplifies many of these challenges. In this paper we present the key methods, knowledge representation requirements, how the underlying data is associated with the smart search and integration and comment on extensibility and applicability to other Earth science application areas.

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

1. Introduction

Increasingly scientists and non-scientists are addressing interesting problems using distributed information products and data resoures from a variety of disciplines.

To ground this work in a relevant and specific example, we describe a use case as follows: 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 atmospheric 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, leading to medium-term climate forcing. 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 (e.g. specific instruments from specific missions and discipline areas) that may not be. 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 current 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; [11, 2, 10]). 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 (for details on the VSTO see later references herein).

We also take advantage of significant experience with ontology packages and data registration from the Geosciences Network (GEON) [6]. Our present ontology developments involved some new material as well as iterations and augmentations of a background domain ontology: the Semantic Web for Earth and Environmental Terminology (SWEET) [7].

In this paper we present the needed paradigm shift, the specific application use case, our methodlogies and details on how we mediate the data integration task before concluding and presenting ongoing work.

2. Changing the paradigm

What is the problem? Scientists only use data from a single instrument because it is difficult to access, process, and understand data from multiple instruments. A typical data query might be:

- "Give me the temperature, pressure, and water vapor from the AIRS instrument from Jan 2005 to Jan 2008"
- "Search for MLS/Aura Level 2, SO2 Slant Column Density from 2/1/2007".

This type of query is typical in present data environments: if you know exactly what you want and do not care about anything else, it is mostly possible to find it. Increasingly, this situation is umcommon and less user needs are being met.

What is a solution? Using a simple process, the work developed and presented here allows data from various sources to be registered in semantically meaningful way (i.e. to an ontology), so that it can be easily accessed and understood across disciplines and diverse data holdings. Scientists (unknowningly) use only the ontology components that relate to their data. A more understandable query might look like:

- "Show all areas in California where sulfur dioxide (SO2) levels were above normal between Jan 2000 and Jan 2007"

This query will pull data from all available sources registered to the ontology and allow seamless data fusion. Because the query is measurement related, scientists do not need to understand the details of the instruments and data types.

3. Use Cases

In keeping with our developed methodology (next section) we have developed several underlying use cases [8] for an initial application of data integration to volcano eruption-atmosphere impacts. A typical expression of 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 being between the troposphere and stratosphere and sensitive to the temperature gradient in the atmosphere.

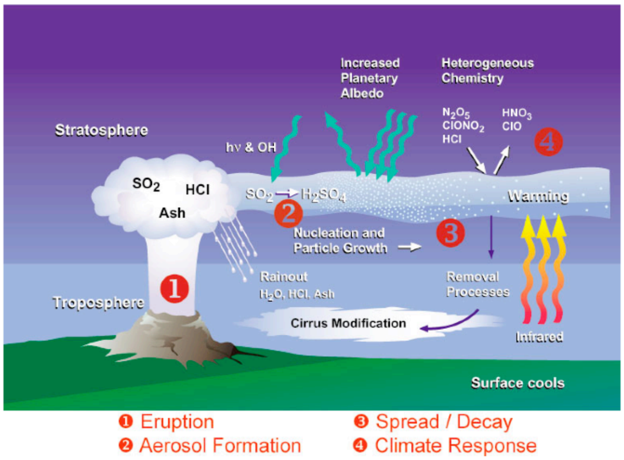

Figure 1. Schematic of the events and processes beginning with a volcanic eruption and leading to the climate/atmospheric response.

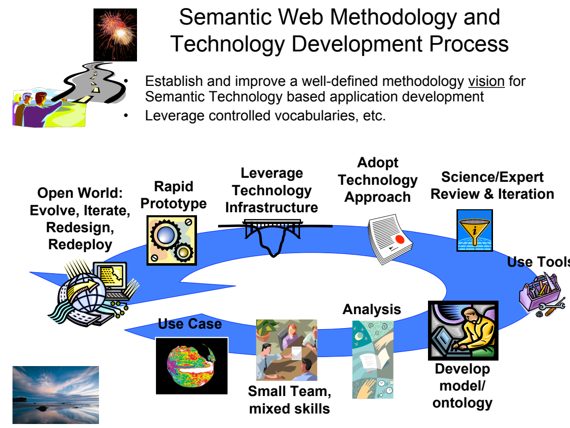

Figure 2. Our developed methodology for application of semantic web to applications (Fox and McGuinness)

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.

4. Methodology

We apply semantic web methodologies in pursuit of the above-mentioned objectives. These methods include the development and elaboration of use cases (user scenarios) with significant science/subject matter and data expert involvement. 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 from IHMC; http://cmap.ihmc.us/coe) that is capable of reading and writing OWL-based ontologies and provides OWL predicate assistance when users are adding relations between concepts (e.g. is-a, has, disjoint, etc.). The starting points going into these workshops and their nominal end-points (although not the end product) are recorded in concept map form as intermediate artifacts.

Our data integration effort depends on machine processable specifications of the science terms that are used in the disciplines of interest. Based on our starting points for ontologies as well as those we are re-suing, we have identified specific ontology modules that needed construction in the areas of volcanoes, plate tectonics, atmosphere, and climate. We bring together a small group of 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.

The overall methodology we employ is denoted schematically in Fig. 2 beginning with use case(s), the small team (eluded to above), analysis of the use case, modeling and ontology development using available tools and then expert review. After that we adopt suitable technical approach(es) and specific technical infrastructure (leveraging existing work to the extent possible) and rapidly develop and deploy something that can be tested, evaluated and then iterated upon; revisiting the full methodology cycle.

5. Mediating the Data Integration

The result of applying the methodology up to and including iteration on the science expert review resulted in the application specific ontology packages indicated in Figure 4. We show a portion of the key important concepts and relations for detection and attribution of the present use case in the concept map in Fig. 3.

In relation to climate effects there are atmosphere layers and atmosphere layer boundaries that are part of climate. These concepts have subclasses such as

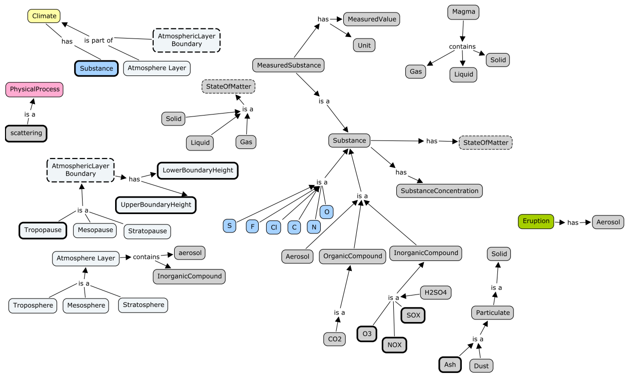

Figure 3. An excerpt of the most recent version of the atmosphere/climate concept map with specific integration concepts that are motivated by the use case (detection and attribution).

troposphere and tropopause (respectively) and each has properties such as lower and upper boundary heights. 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 layers have primary substances, 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. 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 point for our application.

Ultimately, it is compounds such as SO_2 concentration that are measured and are were the answers (data) to queries such as that given in Section 2 are to be found.

5.1. Packaging the Ontology and Services

For the ontology development we utilize a modular approach (which is considered best-practice in the semantic web methodology community). Thus as we developed the classes and sub-classes in the volcano, plate tectonics and atmosphere ontologies, we associated them with one of the ontology modules indicated in the Fig. 4, which is a schematic of our approach the present application and indicates how we leverage/ import many other ontologies.

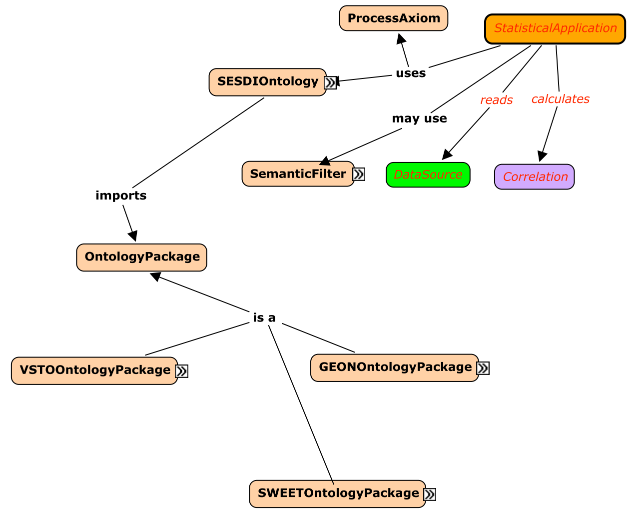

Figure 4. High-level concept map indicating how a statistical application makes use of the data, any filters (e.g. restrict attention to geochemical measurements or to volcanoes, see later for details), and the underlying knowledge base, i.e. ontologies.

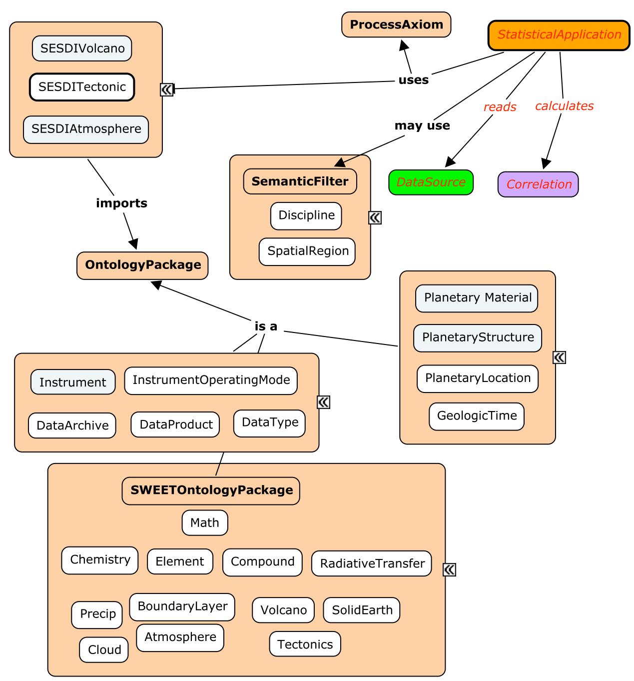

Figure 5. Expanded view of Fig. 4. See text for details.

Note in Fig. 5, the ontology package directly connected to the application contains only the relating concepts needed for the data integration and imports fundamental terms and concepts from SWEET, VSTO (for instruments, data, etc.) and GEON (for planetary specific concepts). VSTO has a flexible and re-usable ontology that we have extended with instrument subclasses, instances and measured parameters relevant to the application areas of volcano and climate. We also import modules from solid earth concepts from the GEON project and substantial components from the newly modularized α-version 2.0 of SWEET (Raskin, private communication; soon to be published). The detailed discussion of each of the imported modules will be presented in a later paper.

5.2. Data Registration

As noted above, the next key element in semantic data integration is associating the semantic terminology with the relevant data sources. One of the Geosciences Network [6] project's contributions has been a three step view of registering data [12] 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 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.

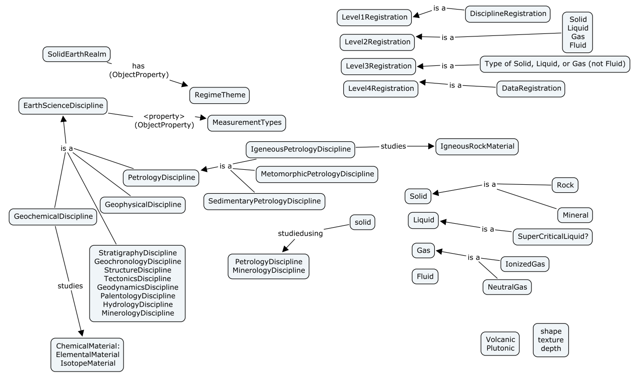

Figure 6. Concept map (ontology) for data registration levels 1, 2 (here called level 2 and 3) and 3 (here called level 4) of volcanic data.

Our adaptation of this procedure has been implemented in a new desktop application called SEDRE - Semantically-Enabled Data Registration Engine. Each of the three stages is driven by casting the use case stated above in terms of an actual data provider wishing to register their data. Most importantly, the way data prodivers think about their data, i.e. how to classify it according to the three levels, differs between disciplines (e.g. volcano geochemistry and atmospheric chemistry in our example here). To capture these differences within the SEDRE application we model the discipline specific registration using a concept modeling approach which we use to create in declaritve form, a data registration ontology (see Fig. 6). Thus, instead of forcing a common and unfamiliar registration method, we may accommodate the differences within one application. The concept map in Fig. 6 highlights the way solid-earth researchers think about classifying their data. E.g. the first level could be Petrology (or a sub-discipline of it), then Rock, and then the type of Rock, and then, the specifics of the data measurement. In our example, it is Geochemical, Gas (NeutralGas), and then the data for SO2. We also use an ontology for atmospheric data set (e.g. chemistry) registration (not shown here).

We have registered a series of volcanic geochemistry and atmospheric chemistry data sets. An example of the volcano data registration is shown in Figs. 7-8 for a dataset in spreadsheet form from the Kilauea east rift zone (Hawaii Volcano Observatory). We omit the example of atmospheric data registration here but examples are MLS (Microwave Limb Sounder) on NASA AURA mission (level 2; swath) and the ESA SCIAMACHY (SCanning Imaging Absorption SpectroMeter for Atmospheric CHartographY, also level 2) mission

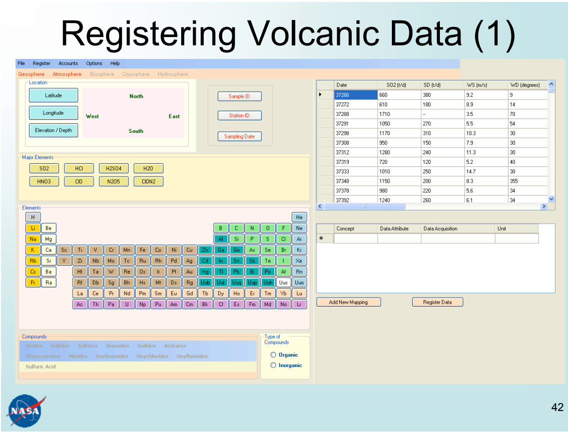

Figure 7. Screen capture of initial stage of volcano geochemical dataset registration.

data, most often in HDF (Hierarchical Data Format).

Figure 7 shows a screen shot of the prepatory phase where the geosphere registration is selected and a dataset in spreadsheet form is loaded into the upper right pane. The upper left features discovery level information such as latitude and longitude. The next set of panes on the left have major elements, the periodic table and compound classes.

Figure 8 shows shows an intermediate screen shot registration where the SO2 entry has been selected from the major elements pane and associated with the relevant column in the spreadsheet table, also SO2. The entries are visually recorded as the user selects them in the middle right pane. In addition, the fact that this is a measured quantity (e.g. as distinct from proxy, modeled, or inferred) and the units (t/d = tonnes/day = 1000kg/day) are recorded by the user. This greatly assists in later use of the data, especially for data integration and data fusion. Finally, we note that in this case since the measurements are in-situ, the latitude and longitude are not explicit in the dataset but are to be inferred from the volcano name.

5.3. Leveraging the VSTO data framework

The Virtual Solar-Terrestrial Observatory [11, 2, 10] has developed a production semantic data framework in support of the solar, solar-terrestrial and space physics observational communitities. Fig. 9 is schematic of the high-level organization of the VSTO framework applied to the present use case. This figure displays how we have been able to immediately

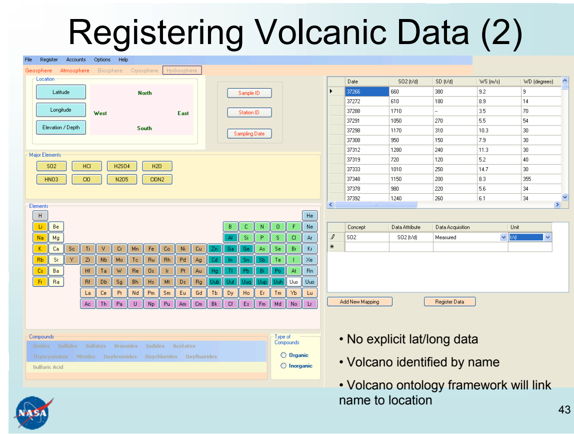

Figure 8. Screen capture of intermediate stage of volcano geochemical dataset registration.

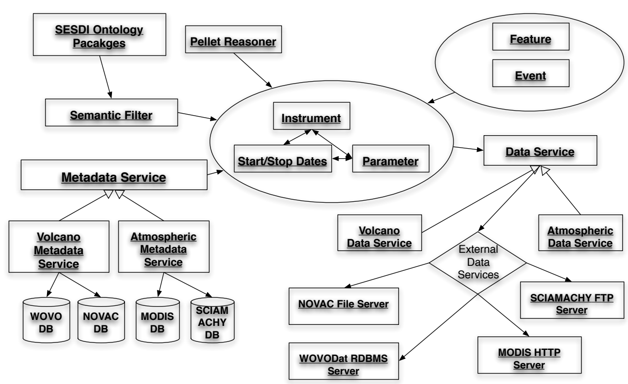

Figure 9. Schematic of the VSTO data framework applied to the current use case. Note the loading of the SESDI ontology and leveraging existing distributed data (and where needed, metadata) sources.

leverage the VSTO framework (this diagram is a direct copy of the VSTO implementation with the solar-terrestrial-specific ontology and data sources replaced by the appropriate volcano and atmospheric ontologies and data/ catalog sources). The primary addition is the Feature and Event classes, which are required to represent volcanoes, and eruptions, for example. 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. Note that we also re-use packages from the VSTO ontology, populated with instruments, etc. specific to the domain application (as noted earlier).

6. Discussion and Conclusion

We have presented and discussed the important elements required in addressing the needs of integrating inter-disciplinary data from diverse sources. We have outlined and given examples of our semantic web methodology, tools and processes we use, and the ontologies we have developed and re-used. They key element of registering data to the underlying semantics is now reaching a much more mature stage of development and is undergoing user testing. In future papers we will present evaluations of this (and other) use of our tools.

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

- We are substantial reducing the number and extent of ontologies we need to develop due to the modularizing/ packaging approach.
- We are significantly re-using ontologies developed by others.
- We are leveraging implementations of semantic infrastructure.
- 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 for the larger atmosphere ontology effort.
- Since the concepts and relations specific to the interdisciplinary domains are loaded selectively (by design) we believe that the ontology package approach is applicable among numerous disciplines and applications. In the latter phase of the current project, we will test this assertion by integrating data from solar radiation and climate response, and also by the three-way test of volcano and solar forcings signatures on the atmosphere.

Finally we move forward in this work, we plan to make the SEDRE application available for data providers to download and use to register their datasets. For implementing the use cases, in addition to utilizing the VSTO framework, we plan to work towards utilizing the the DIA engine developed by the GEON project [13, 14, 15] which 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, and/or scientists to share their resources (e.g., geochemical data, filtering services, etc.). DIA is designed to work with the three level registration procedure we have adopted.

7. 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. Particular thanks go to Rob Sherwood, Karen Moe and Francis Lindsay.

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 Scien-

tific 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, USGS Scientific Investigations Report 2007-5199, 43-46.

[2] 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. and AI Magazine, vol. 29, no. 1, 65-76.

[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, Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience, Computers and Geosciences, in press.

[11] P. Fox, D. McGuinness, R. Raskin, A. K. Sinha 2007, A Volcano Erupts: Semantically Mediated Integration of Heterogeneous Volcanic and Atmospheric Data, ACM Proceedings of the CyberInfrastructure: Information Management in eScience (CIMS).

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

[13] 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.

[14] 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.

[15] 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.