Current and future uses of OWL for Earth and Space science data frameworks: successes and limitations

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

[1] McGuinness Associates, Stanford, CA 94305 USA
[2] Stanford University, Stanford, CA 94305 USA
[3] High Altitude Observatory, National Center for Atmospheric Research, Boulder, CO 80307 USA
[4] Scientific Computing Division, National Center for Atmospheric Research, Boulder, CO 80307 USA

{dlm, jbenedict}@mcguinnessassociates.com
{pfox, luca, pwest, jgarcia}@ucar.edu

Abstract. Based on almost three years of experience in developing and deploying scientific data frameworks built using semantic technologies, we now have a production virtual observatory in operation, serving two broad communities: solar physics and terrestrial upper atmospheric physics. Within this application, a data framework provides online location, retrieval, and analysis services to a variety of heterogeneous scientific data sources that are often highly distributed over the internet. In this paper, we describe selected current and planned uses of OWL-DL, related tools, and our deployment. We describe some successes and limitations we have found to date using OWL-based technologies, especially tool support. We also indicate the important components we require from a robust technical infrastructure as we move forward with expanding the functionality of the frameworks. This expansion includes additional semantic representation and reasoning/query services as well as broadening the scope of our scientific disciplines.

Keywords: Virtual Observatory, Semantic Integration, OWL, Reasoning, Semantic Query, Scientific Data, Geosciences, Solar-terrestrial physics, volcanoes, climate, applications,

1 Introduction

There is a growing need to find, access, and use large amounts of distributed interdisciplinary scientific data. Solutions to address this need in the form of integrated data systems, distributed data frameworks (DFs) and Virtual Observatories (VOs) are also proliferating. VOs present the access point for distributed resources containing large volumes of scientific observational data, theoretical models, and analysis programs and results from a broad range of disciplines. Our recent work,

spanning a three year period on two scientific data-intensive projects (funded by NSF and NASA) provides the setting from which we report our findings. VOs intend to make all resources appear to be both local and integrated; our approach to this goal is to use semantic technologies.

Our initial science domain areas were solar, solar-terrestrial, and space physics. These domain areas required a balance of observational data and theoretical models to combine many data sources with various origins. Previously, even the experienced researcher needed to know a significant amount about the instruments and models as well as arcane and obscure related information such as acronyms and numerical codes for instruments operating in particular periods and modes. We have built a semantically enabled platform that supports scientific data integration. The primary project we are reporting on here integrates data between volcano events and local and regional climate settings, and then enables search and inference across the integrated interdisciplinary collection. One requirement we had was to move the data search and access for such integration from an instrument-based approach to a measurement based approach. For example many different instruments in varying locations may measure SiO_2 both in rock/mineral samples and in the atmosphere. At present, users have to know which instruments made the right type of measurements and they have to navigate the particular peculiarities of each set of data holdings. For example, the names of an otherwise identical measurement may be different between databases. The units of measure may be different and not well documented. Further, the associated metadata and cataloging may not make it possible to find certain measurements. To allow a user to search by measurement requires establishing the relations between instruments and what they measure and vice-versa. Thus, a data framework is required that represents and relates important concepts and processes (in the application area) and precise relationships are known and encoded. The framework also needs to link these concepts, processes and relationships to the underlying data. One end use of a semantic framework is to bring diverse data into an application, perhaps statistical, which could be used to evaluate the hypothesis of a connection between volcano emissions and effects on atmospheric air quality.

The key to achieving the VO and measurement-based data integration vision is in providing users (humans and agents) with tools and services that help them to understand what the data is describing, how the data relates to data possibly in another topic area, how the data was collected, and the implicit and explicit underlying assumptions. We refer the reader to previous work on the interdisciplinary Virtual Solar-Terrestrial Observatory (VSTO) for more about the architecture and applications [www.vsto.org, Fox, McGuinness, et al, 2006, McGuinness, Fox et al. 2006]. In this paper we report on our latest experience with relevant OWL-based ontologies, describe how we are leveraging existing background domain ontologies, and provide an overview of how we generate our own ontologies covering the required subject areas. Further we report on selected critical surrounding tools and infrastructure required to build operational semantic web applications in our application domains and indicate what functionality we will need from those tools as we move into the future.

2 Use Case Driven Development

In the last year, we have augmented our initial motivating set of VSTO use cases. In general form the original use cases are noted in templates/examples 1 and 2 and the newer use cases present more generalized and science-relevant patterns and are noted in templates/examples 3 to 6.

Template 1: Plot the values of Parameter X as taken by instrument description or instance Y subject to constraint Z during the time period W in style S. Example 1: Plot the observed/measured Neutral Temperature (Parameter) looking in the vertical direction for Millstone Hill Fabry-Perot interferometer (Instrument) from January 2000 to August 2000 (Temporal Domain) 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 images of the solar corona during a recent observation period.

Template 3 Find data for parameter X constrained by Y during times described by Z. Example 3: Find data, which represents the state of the neutral atmosphere anywhere above 100km and toward the Arctic circle (above 45N) at any time of high geomagnetic activity.

Template 4: Assemble a visual representation of a sequence of images X over a time period Y: Example 4: Create a movie of the white light solar corona during the whole-Sun campaign month in 2005.

Template 5: Infer data representing a state of one physical domain X that changes in response to an external event Y from another physical setting Z. Example 5: Find and plot/animate data that represents the terrestrial ionospheric effects of a geo-effective solar storm.

Template 6: Expose semantically enabled, smart data query services via a web services interface allowing composite query formation in arbitrary workflow order. Example 6: 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.

We followed the same methodology we used previously when building our ontologies driven by uses cases [Fox et al. 2006, McGuinness et al. 2006, McGuinness et al 2007]. In brief, this meant extracting the key vocabularies to determine classes, sub-classes, associations and initial key properties as well as underlying data sources and end use requirements for the returned data. The expanded use cases did not lead us to expand the science coverage much; they resulted in the need to integrate across domain areas. However, we did need to re-examine the simplifications we had initially put in place in the class and property structure of the ontology. We needed to add the event, process and phenomena concept categories, which previously had not been required. However, these additions did not alter our original class and property structure since the two sets were orthogonal, i.e. each distinct upper-level class element was faceted and thus modular.

Figure 1 represents the high-level interaction view of how selections and services are combined in the VSTO data framework. Based on three of the abstract level classes from the VSTO ontology (upper left) and semantic filters, together with reasoning, the central selection procedure has been integrated across a variety of

previous data workflows down to the basic combination of instrument, date/time and parameter. This was a significant and unexpected outcome of the ontology development and allowed one portal and set of web services to provide access to data holdings ranging from solar physics images to incoherent scatter radar data as a function of time and altitude. A substantial portion of the VSTO ontology addresses the need to both retrieve metadata from external sources as well as the data itself. The metadata concerns both classes and instances not encoded in the ontology. Our data services are in essence a semantic abstraction of the previous data services and these services allow users to obtain the data that is essential for carrying our scientific investigations.

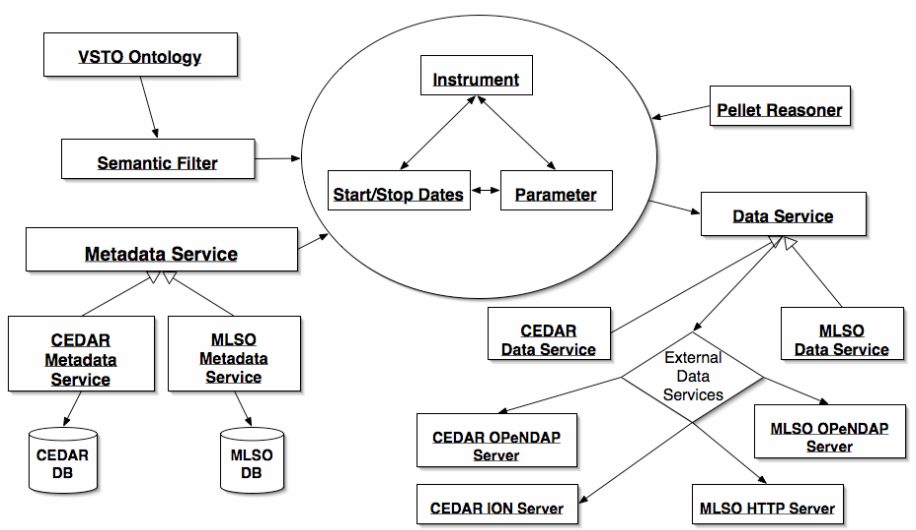

Figure 1. Relation of semantics, data selection workflow and external services for the VSTO production portal based on first two use cases.

3 Developing and Encoding the Ontologies

We used the newer use cases to drive the ontology expansion. We focused first on expanding the instrument ontology. One challenge for integration of scientific data taken from multiple instruments is in understanding the conditions under which the data was collected. It is important to collect not only the instrument (along with its geographic location) but also its operating modes and settings. Scientists who need to interpret data may need to know how an instrument is being used – i.e., using a spectrometer as a photometer. (The Davis Antarctica Spectrometer is a spectrophotometer and thus has the capability to observe data that other photometers may collect). An advanced notion is capturing the assumptions embedded in the experiment in which the data was collected and potentially the goal of the experiment.

In Figure 2 the descriptions of the classes relevant to our examples are:

- Instrument: A device that measures a physical phenomenon or parameter.

- OpticalInstrument: An instrument that utilizes optical elements, i.e. passing photons (light) through the system elements.
- Photometer: A transducer capable of accepting an optical signal and producing an electrical signal containing the same information as in the optical signal. The two main types of semiconductor photodetectors are the photodiode (PD) and the avalanche photodiode (APD).
- SingleChannelPhotometer: Photometer that samples with one specified restricted wavelength/frequency range.
- Spectrometer: An optical instrument used to measure properties of light over a specific portion of the electromagnetic spectrum. A spectrometer is used in spectroscopy for producing spectral lines and measuring their wavelengths and intensities. Spectrometer is a term applied to instruments that operate over a wide range of wavelengths; gamma rays and X-rays into the far infrared.
- Spectrophotometer: A spectrometer that measures light intensity. (It can also record the polarization state (which includes intensity)).

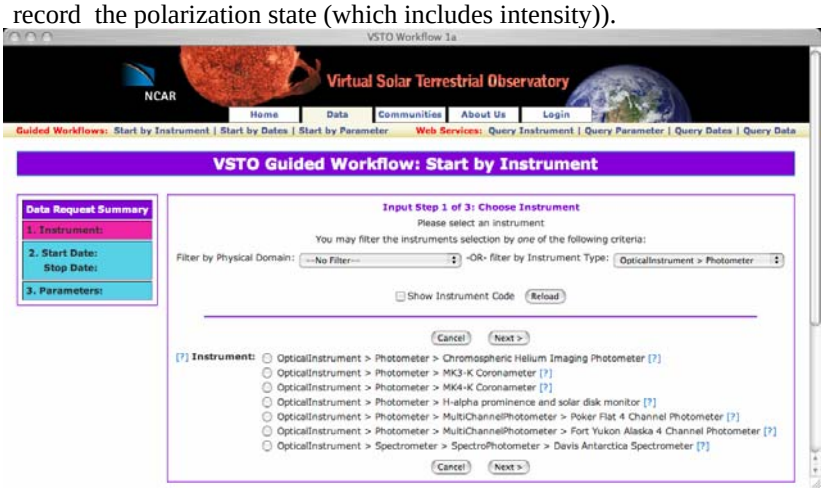

Figure 2. Portion of VSTO ontology 1.0 indicating that with certain properties a Spectrophotometer can act as a photometer and that filtering instrument selection will include the spectrophotometer (when applicable) and that instrument choices will be available that previously were not.

4 Data integration across discipline boundaries

Another need in science disciplines is to provide smarter software for integrating data. Our integration use cases need to integrate data across discipline boundaries, in pursuit of solving problems that today take months and years to assemble, explore hypotheses, and validate conclusions. One motivating example is the study of the local and regional effects on climate of volcanic activity. The appearance of episodic perturbations in the climate record on a global scale correspondence with the occurrence of medium and large volcanic eruptions (e.g. El Chicon in 1982 and Mt. Pinatubo in 1991) is well known [see earthobservatory.nasa.gov/Study/Volcano].

We incorporate this discussion of data integration since it drives a particular method of developing the required ontologies as well as differing applications needing to be developed. In the virtual observatory example data is returned via a web portal or web service. The system response is a data product. In contrast, for the volcano-climate study, there is a need to embed a semantic representation (or reasoning) directly within the user's application, i.e. terms (classes and properties) and relations need to be returned to the user application and reasoned with before suitable data is identified and returned. Later, once these reasoning services are developed and generalized, we expect to build a set of (web) services on top of the existing ones provided on the server side framework.

To build the set of required ontologies, we utilized the same small teams [Fox et al. 2006, McGuinness et al. 2006, McGuinness et al. 2007] as in the virtual observatory process but focused on more generality: We needed to model a broad set of concepts and relations in volcanic settings with an emphasis on volcanic phenomenon that lead to atmospheric perturbations. When working with domain experts, we have found that working with a visual representation of the ontology (especially portions of it) is by far the best method of knowledge capture and iteration. We found the visual representations to surpass plain English and any form of OWL representation. We utilize the concept-mapping (CMAP) tool from IHMC (http://cmap.ihmc.us) for this purpose. At later stages, we translate the concept map into UML and OWL-DL for application use.

Figure 3 shows the high level concepts for earth settings including volcanoes and related features. In this figure we see that a volcano is a subclass of a volcanic system, which has properties such as name, shape, environment, and climate (to name a few). What became apparent in connecting to underlying data sources was that the tectonic setting and its attributes were essential to capture to consistently represent the volcano and its phenomena. As a result, we initiated a related ontology modeling exercise (see an excerpt from this effort in Figure 3. Perhaps the important aspect of this figure is that (toward the bottom) it displays how measurements of certain phenomena (e.g. eruption is a volcanic phenomena which has measurements such as: mass flux, seismic energy, etc. to characterize it) connect to other concepts and relations in the ontology. It is these measurements that directly connect to elements in the databases we seek to exploit for data integration. Figure 4 shows the modular approach being taken in this project and related projects (e.g. GEON; the Geosciences Network, www.geongrid.org, which also has generated modular packages of ontologies complementing this work). It shows imported packages from SWEET (Semantic Web for Earth and Environmental Terminology and OWL-Time [Hobbs and Pan 2004] leading to an aggregate set of concepts and relations for Planetary Material. (Courtesy of K. Sinha; private communication).

The next phase of the project involves following the same knowledge acquisition process used previously for obtaining critical classes and properties in the atmosphere / local climate domain. This domain is well-covered in the SWEET ontology, thus we will attempt to reuse as much as possible. We have populated a CMAP using the relevant SWEET classes and properties and this will be used our subject matter experts, who will be driven by our use cases to augment, prune, and refine as necessary.

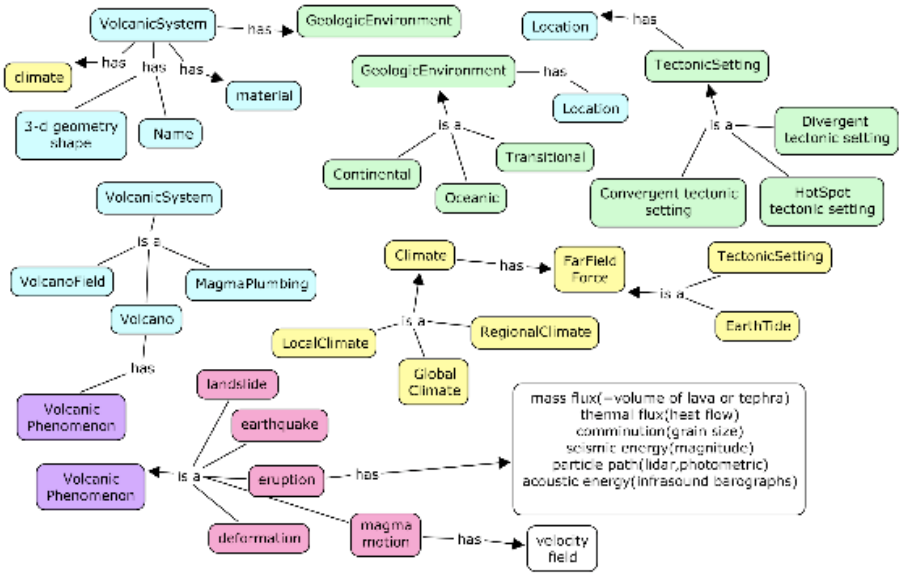

Figure 3. An excerpt from the volcano ontology model using CMAP notation [McGuinness et.al, 2006].

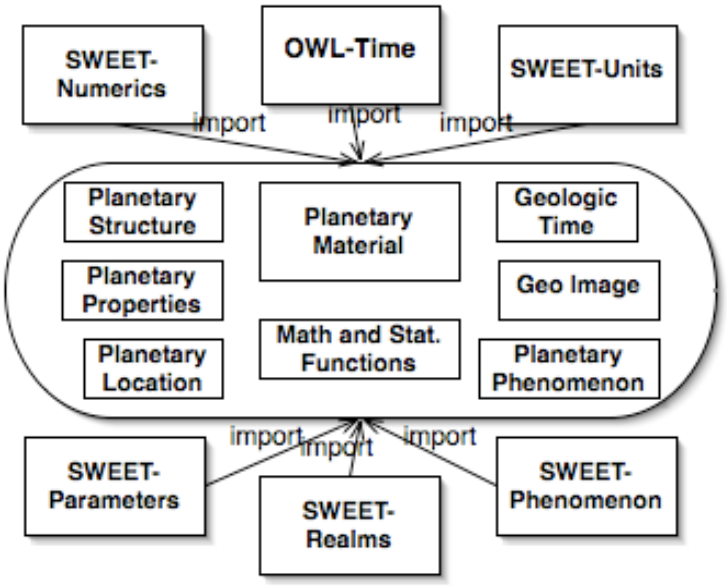

Figure 4. Ontology package architecture for data integration in Earth Sciences.

To achieve the science goals, we need to connect data sources to the overlying knowledge framework as discussed above. We immediately recognized that the entire

VSTO ontology covering instruments, observatories, data archives, and data products was directly reusable in this project. The only additions we needed were some straight forward additions of instruments appropriate for volcano research. Figure 5 shows such an extension to the Spectrometer class to add Mass Spectrometers of various types and the instances. All properties that we had added for the Spectrometer class for VSTO (solar and solar-terrestrial physics) were applicable to inherit for the Mass Spectrometers used in compositional analysis for volcanoes.

Lastly, we needed to identify the quantities (Parameters) that were measured by these instruments (not shown). We found that some of the parameters were already encoded in the SWEET ontology and many were also in the GEON ontology. We are presently registering a number of volcanic databases with this portion of the ontology in preparation for the data integration application.

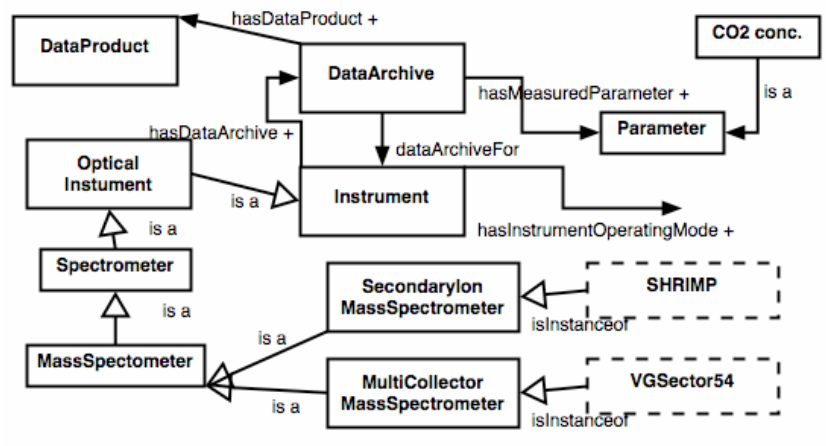

Figure 5: Augmentation of VSTO instrument class structure to include classes of Mass Spectrometers in use for measuring rock properties in volcanic settings (this is a partial list of all instruments to be used in this application).

5 Reliance on Semantic Technology Tools and Documentation

We believe that a critical aspect to our success with using, deploying, and disseminating our semantic web-based applications is availability of tools and documentation. We did a careful selection of our core team and core tools and we believe that enabled us to generate prototypes quickly and to create an extensible infrastructure. Internally, we heavily leveraged the Protégé[1] and Swoop[2] editors and the Pellet reasoner. We also leveraged a number of the Protégé plug-ins, the most critical of which was the one that generated java classes, since java compatibility was essential. We also leverage species validators. As we brought our internal team up to speed, we relied heavily on the OWL Overview, Guide, and Reference Manuals in

[1] http://protégé.stanford.edu/
[2] http://www.mindswap.org/2004/SWOOP

addition to Ontology 101 [Noy, McGuinness, 2001] and Ontologies come of Age [McGuinness, 2003] papers. Additionally, before we went into any of our knowledge acquisition sessions, we asked attendees to read the Ontologies come of Age paper and glance at the OWL Overview and the Ontology 101 paper. We also looked for controlled vocabularies and ontologies that would be considered reasonable starting points and we came prepared with them in the CMAP tools as well as sometimes in OWL tools (again SWOOP and Protégé). We found that the foundational, accessible papers were critical in order to give our domain experts some idea about what ontologies were and how they might be used. We also found that tools like CMAP that have very low barriers to entry are good tools for brain storming sessions such as knowledge acquisition meetings. While, of course CMAP provides enough flexibility for users to hang themselves (in that it allows any label on any link between any nodes, i.e. they provide arbitrary semantics and no validation methods), they can be used effectively to gather controlled vocabulary terms, and, with a facilitator, they can be used quite effectively to gather more formal specifications.

Further, we believe that we have just scratched the surface for our outreach effort. We believe that the documentation we relied on to bring people up to speed with simple discussions and simple examples will be even more critical as we expand our efforts into broader science domains. Our goal is to do less hands-on work personally as we expand our project reach, thus we believe documentation and tools will become more critical. Some tools we are also just starting to use that we also think will be critical include explanation environments, such as Inference Web [McGuinness, et. al, 2004] ontology evolution environments, such as Chimaera [McGuinness, et. al, 2000], and ontology search tools, such as SWOOGLE [Finin, et. al, 2005]. We also believe documentation on the life cycle point and progression of the tools and underlying language(s) to be an important component for adoption.

7 Summary

We designed, implemented, and deployed a semantic data framework for virtual observatories covering content in solar and solar-terrestrial physics. We have taken this deployed framework and expanded it to support data integration across volcanic and regional climate settings. Our ontology-enhanced services and tools provide retrieval, analysis, and plotting support. We have found that the general framework is robust and extensible. We believe documentation on the tools and simple examples to be critical to broad adoption. We have also found that editor, reasoning, and environmental tools to be increasingly critical to adoption. Once users become dependent on these environments, we are finding it increasingly important for them to have continuing support with respect to life cycle maintenance of tools and also for the tool and language developers to provide migration path support if updates are made.

Acknowledgments. The authors acknowledge funding from the National Science Foundation, SEI+II program under award 0431153 and NASA/ACCESS and NASA/ESTO under award AIST-QRS-06-0016.

References

1. Tim Finin, Li Ding, Rong Pan, Anupam Joshi, Pranam Kolari, Akshay Java, and Yun Peng. Swoogle: Search for Knowledge on the Semantic Web. Proceedings of the American Association for Artificial Intelligence Conference (AAAI05). July 29, 2005.
2. Peter Fox, Deborah L. McGuinness, Don Middleton, Luca Cinquini, J. Anthony Darnell, Jose Garcia, Patrick West, James Benedict, and Stan Solomon. Semantically-Enabled Large-Scale Science Data Repositories, in Lect. Notes in Comp. Sci., ed. Cruz et al., vol. 4273, pp. 792-805, Springer-Verlag, Berlin, 2006
3. Peter Fox, Deborah L. McGuinness, Rob Raskin, A. Krishna Sinha. Semantically-Enabled Scientific Data Integration. U.S.Geological Survey Scientific Investigations Report 2006-5201, p. (Geoinformatics 2006)
4. Jerry R. Hobbs and Feng Pan. 2004. An Ontology of Time for the Semantic Web. ACM Transactions on Asian Language Processing (TALIP): Special issue on Temporal Information Processing, Vol. 3, No. 1, March 2004, pp. 66-85.
5. Deborah L. McGuinness. ``Ontologies Come of Age ". In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2003.
6. Deborah L. McGuinness, Richard Fikes, James Rice, and Steve Wilder. *An Environment for Merging and Testing Large Ontologies.* In the Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado, USA. April 12-15, 2000.
7. Deborah L. McGuinness and Paulo Pinheiro da Silva. Explaining Answers from the Semantic Web: The Inference Web Approach. Web Semantics: Science, Services and Agents on the World Wide Web Special issue: International Semantic Web Conference 2003 - Edited by K.Sycara and J.Mylopoulis. Volume 1, Issue 4. Journal published Fall, 2004. http://www.websemanticsjournal.org/ps/pub/2004-22
8. Deborah L. McGuinness, Peter Fox, Luca Cinquini, J. Anthony Darnell, Patrick West, James L. Benedict, Jose Garcia, and Don Middleton. Ontology-Enabled Virtual Observatories: Semantic Integration in Practice. CEUR Workshop Proceedings, vol. 216, online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-216/submission_14.pdf
9. Deborah L. McGuinness, A. Krishna Sinha, Peter Fox, Rob Raskin, Grank Heiken, Calvin Barnes, Ken Wohletz, Dina Venezky, Kai Lin. Towards a Reference Volcano Ontology for Semantic Scientific Data Integration. Eos Trans. AGU 87(36), Jt. Assem. Suppl., Abstract IN42A-03.
10. Deborah L. McGuinness, Peter Fox, and Don Middleton. Solar-Terrestrial Ontologies. . U.S.Geological Survey Scientific Investigations Report 2006-5201, p. (Geoinformatics 2006).
11. Deborah McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict & Don Middleton, The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In Proceedings of Innov. Applic. Artif. Intell., in press. 2007.
12. Natalya F. Noy and Deborah L. McGuinness. ``Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
13. Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: a practical OWL-DL reasoner. Submitted to Journal of Web Semantics. *http://www.mindswap.org/papers/PelletJWS.pdf*