Semantically-Enabled Virtual Observatories

Deborah McGuinness[12], Peter Fox[3], Luca Cinquini[4], Patrick West[3],
James Benedict[2], J. Anthony Darnell[3], Jose Garcia[3], and Don Middleton[4]

[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, tdarnell, jgarcia, don}@ucar.edu

Abstract. We are developing a semantic data framework for virtual observatories. A Virtual Observatory provides online location, retrieval, and analysis services to a variety of heterogeneous scientific data sources. We employ semantic technologies to integrate data and provide "intelligent" services such as ontology-enhanced search, analysis, and data visualization. Our specific initial deployments are in the field of solar-terrestrial physics where we target atmospheric and solar researchers as end users. In this paper, we describe our general use case, our approach using OWL-DL and related tools, and our initial deployment. We describe what we have found as benefits and challenges using OWL-based semantic technologies in our efforts building an operational system. Our system is deployed in two scientific data collections with community usage migration starting now.

Keywords: Virtual Observatory, Semantic Integration, Scientific Data, Solar-terrestrial physics, applications,

1 Introduction

Semantic technologies are a potential key enabler for Virtual Observatories (VOs) to effectively meet the challenges of modern scientific data discovery, access and use. VOs are distributed resources that may contain vast amounts of scientific observational data, theoretical models, and analysis programs and results from a broad range of disciplines. While we are concerned with Virtual Observatories in general, our initial science domain areas are solar, solar-terrestrial, and space physics. These domain areas require a balance of observational data and theoretical models to make effective progress. They require a combination of many data sources with various origins typically requiring much from even the experienced researcher. Users need to know a significant amount about the instruments and models as well as arcane and

obscure related information such as acronyms for instruments operating in particular periods and modes. Additionally, since many of the data collections are increasingly growing in volume and complexity, the task of truly making them a research resource that is easy to find, access, compare and utilize is a significant challenge to discipline researchers who often cannot keep up with all of the updates, and thus will not find key data without infrastructural support. The datasets can be highly interdisciplinary as well as complex. They provide a good initial focus for virtual observatory work since the datasets are of significant scientific value to a set of researchers and capture many, if not all, of the challenges inherent in complex, diverse scientific data.

The virtual observatory (VO) vision includes a distributed, virtual, ubiquitous, semantically integrated scientific repository where scientists (and possibly lay people) can access data. The data repository is intended to appear to be local. The tools and services should make it easy for users to access and use the data they want. Additionally, tools and services should support users in helping them understand the data, its embedded assumptions, and any inherent uncertainties in a discipline-specific context. The key to achieving the VO 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 (and topic area) relates to other data (and other topic areas), how the data was collected, and what assumptions are being used. These problems are an ideal match for semantic technologies.

We utilize semantic technologies to create the interdisciplinary Virtual Solar-Terrestrial Observatory [VSTO, Fox, McGuinness, et al, 2006]). This requires a higher level of semantic interoperability than was previously required by most (if not all) distributed data systems or discipline specific virtual observatories. We use semantic technologies to bridge the disciplines, supporting identification and use of previously unknown data sources measured by instruments or calculated by models in a simple and scalable way. We leverage existing background domain ontologies [SWEET] and generated our own ontologies in OWL covering the required subject areas. We leverage the precise formal definitions of the terms in supporting semantic search and interoperability.

2 Use Case Driven Development

Our general use case is of the form “Find values for a parameter and plot them in a manner that makes sense for the data”. Variations on this theme include finding data (and parameters, instruments, and observatories) according to topic areas, time periods, target observational areas, etc., of interest to the researcher. The first use cases we developed targeted solar researchers interested in solar activity and the state of the neutral terrestrial upper atmosphere which controls the upper level winds and global circulation patterns and interacts with the ionized portion of the atmosphere. Relevant information needs to be gathered from data from multiple observatories (under different organization’s control), using different instruments in various operating modes and may be complemented by models without the user needing to

know all the details and names of the data sources. This need translates into an evaluation metric: do users find and use data they could not find before? This metric is evaluated using session statistics and analyzing the resulting selections and queries. We will describe how our use cases contributed to our ontology and semantic web architecture requirements.

A restated form of our query is: "Plot values of a particular parameter as recorded by a particular instrument subject to certain constraints in a particular time period, in a manner that makes sense for the data." An instantiation of this pattern that may be asked of our implemented system is: "Plot the observed/measured Neutral Temperature as recorded by the Millstone Hill Fabry-Perot interferometer while looking in the vertical direction during January 2000 in a way that makes sense for the data." The current production portal (www.vsto.org) implements this use case and leads to a graphical representation of the temperature as a function of time.

This use case serves as a prototypical example for our target scientific community, that if answered will help the scientists do their research more efficiently and in a more collaborative manner. Our goal from a semantic web perspective is to demonstrate the development of the semantic framework for a virtual observatory while leveraging existing data sources and (catalog and plotting) services. The anticipated result is a successful return of a graphical representation of the specified data. The workflow for our production release based on an integration of the initial use cases is shown in Fig. 1. The application obtains input from the user (informed by the background ontology and using semantic filters such as the physical domain; solar physics or upper atmospheric physics, or upper-level instrument classes; e. g. all optical instruments, or upper –to-mid-level parameter classes; e.g. all temperature parameters) that infers the observatory, instrument operating modes, type of data, independent variables based an arbitrary user selection from instrument, a time period and parameter(s). Reasoning is used to limit choices at any particular step (and also is used to confirm that the user has permission to access the type of data chosen using authentication information). In Fig. 1 an example of that reasoning is that the selected parameter (from the particular instrument, operating in a specific mode) is a time dependent parameter and thus can only be displayed as a two-dimensional x-y graph. Once a dataset is identified, it is necessary to infer which other parameter in the dataset is the parameter representing time, so that the correct values, units and labels can be shown on the x-axis. In the CEDAR database there is no notion of dependent and independent variables since the vast array of instruments, regions of observation, resulting parameters etc. can often be plotted against many different parameters. Another useful inference is the association of a chosen parameter with a group of related or associated parameters. For example, in this use case, the parameter neutral temperature may be inferred to be associated with other parameters representing other measures of the terrestrial neutral atmosphere, (i.e. neutral density, neutral winds), and also any additional quantities that are recorded at the same time as the measured instrument parameters, (e.g. cloud cover). These quantities may be inferred from related state information as well as the other parameters stored in a dataset file. In the next section we elaborate on our exploitation of more advanced reasoning.

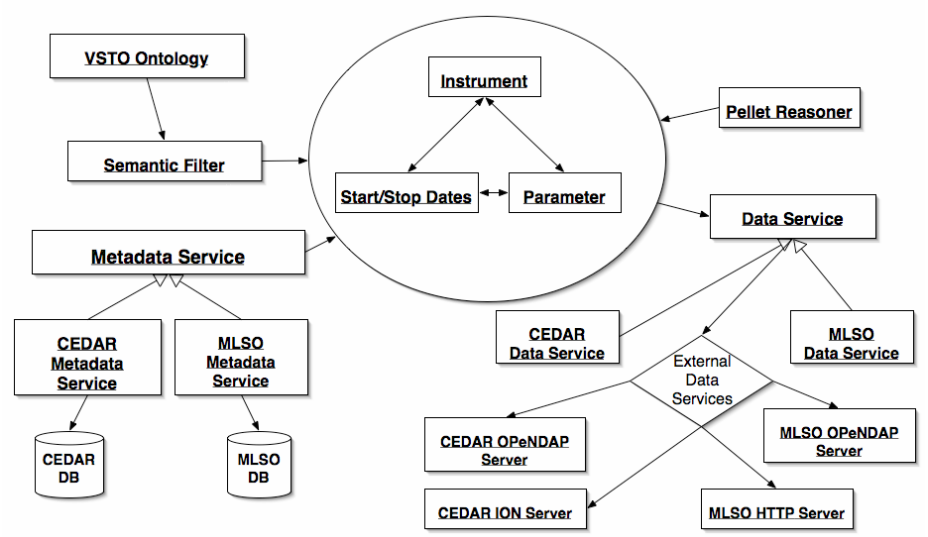

Figure 1. Integrated workflow for VSTO production portal based on first two use cases.

3 Developing and Encoding the VSTO Ontology

We began our ontology development process after carefully analyzing our use cases to look for important classes, instances, and relationships between terms. We analyzed our expected reasoning needs as well and let that drive ontology design and acquisition decisions. We also looked at critical controlled vocabulary starting points that were already included in our base implementations of two existing data services. One such starting point was the controlled vocabulary associated with the CEDAR database, which has a long history in the upper atmospheric and aeronomy communities. For a history of the CEDAR program and the CEDAR database, visit the current website - http://cedarweb.hao.ucar.edu. Data in the CEDAR database was arranged around date of observation and a combined observatory/instrument classification. Within each dataset, a series of tables is encoded in a so-called CEDAR binary format, which holds the parameters. Each observatory/instrument and parameter has a long name, a mnemonic name and a numeric code. An initial pass at the high level classes was made by one domain-literate scientist and one knowledge representation scientist. It became clear quickly that domain expertise was insufficient to develop an extensible and suitably flexible ontology. For example, the domain scientist tended to focus too quickly on properties of classes rather than the class structure and inter-relations between the terms.

In fully developing the ontology, we drew upon both a slightly larger group (5 to 6) and the vocabulary of the use case; the existing vocabulary of CEDAR and wherever possible the terms and concepts in the SWEET ontology. In the case of SWEET, to date there has been limited application to the earth's upper atmosphere (i.e. Realms in SWEET terminology) so we adopted parts of SWEET that applied to our needs and

for the time being, developed our ontology separately from SWEET but keeping in mind that our aim is to merge much of what we develop back into SWEET for broad use. Our goal was to keep our ontology development separate until we believed it was stable and vetted. This also spared us from importing a number of terms at varying levels of detail not directly related to our use cases. We did, however, retain the conceptual composition model of SWEET and reused as many of its terms as possible where applicable with the intent of maximizing our chances of re-integrating our ontology with SWEET – which to date we have not done.

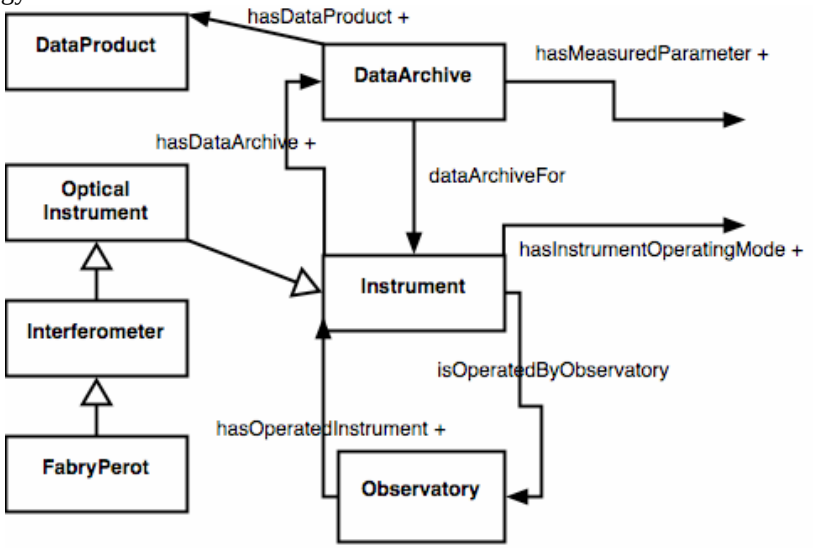

Figure 2. VSTO ontology 0.3 focusing on instruments.

One of the first classes to be discussed in the use case was the concept of an instrument; in this case a Fabry-Perot Interferometer (see description below). One of our contributions both to our domain specific work on VSTO and to general work on virtual observatories is our work on the instrument ontology. We constructed an Instrument class hierarchy (see Fig. 2), including OpticalInstrument, Interferometer and Fabry-Perot Interferometer (as known as FPI, for which the Millstone Hill FPI is an instance of the last class). With each class for the initial prototype we added the minimal set of properties at each level in the class hierarchy. The production release features a more complete but still evolving set of properties across all classes. In the next few paragraphs, we elaborate on a few of the ontology classes in order to give enough background for the impact discussion later. In addition, another use case discussed below introduces the need for inference far beyond the any of the earlier use cases (that tend to map more directly to the classes in the ontology).

In Fig. 2 the descriptions of the classes relevant to our examples follow:

- 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.

- Interferometer: An optical instrument that uses the principle of interference of electromagnetic waves for purposes of measurement.
- Fabry-PerotInterferometer: A particular multiple-beam interferometer. Fabry-Perot interferometers may also be used as spectrometers (i.e. another subclass type of OpticalInstrument with some shared properties are Interferometer but additional ones as well) with high resolution.

In all cases, the class properties are associated with value restrictions, but these are not discussed here. The next important class is the InstrumentOperatingMode (generic description: a configuration which allows the instrument to produce the required signal), which depends on the Instrument and leads to a particular type of physical quantity (parameter; see Fig. 3) being measured and an indication of its domain of applicability and how it should be interpreted.

In practice for the present use case the instrument-operating mode indicates which direction the FPI is pointing, i.e. “vertical” or ``horizontal" - actually 30 or 45 degrees. Knowing these modes is critical for understanding and using the data as different quantities are measured in each mode and geometric projection, i.e. north component of neutral wind has to be calculated correctly depending on the mode.

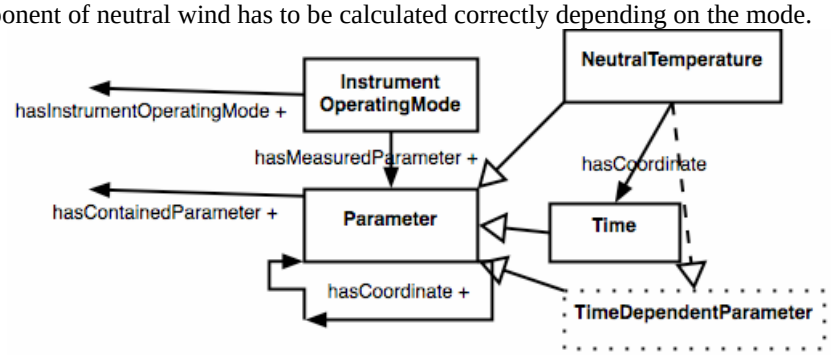

Figure 3. VSTO Ontology 0.3 - focusing on parameters and services

In developing the VSTO ontology we make the connection between the high-level concepts of the ontology classes through to the data files, the data constraints, and the underlying catalogs, and data and plotting with data-related classes.

To satisfy a more advanced use case: “Find data (from CEDAR database), which represents the state of the neutral terrestrial ionosphere anywhere above 100km and toward the Arctic Circle during periods of high geomagnetic activity”, we added a series of properties and additional classes to the ontology. These include PhysicalDomain (domains or realms; which introduces a mapping to the SWEET ontology), PhysicalDomainState (physical state), which includes temperature, pressure, density, winds (for example in the terrestrial neutral atmosphere), and the connection between levels of geomagnetic activity with particular periods of time at which the appropriate instruments are operating. In this case an example of the line of reasoning is: GeoMagneticActivity has the property hasProxyRepresentation and GeophysicalIndex is a ProxyRepresentation (in PhysicalDomain of NeutralAtmosphere). Further, Kp is a GeophysicalIndex, which has the property hasTemporalDomain (whose value is “daily”) and also has the property hasHighThreshold (whose value is 7). Together these inferences allow us to

determine a set of dates/times when the geophysical index Kp is greater than or equal to seven as well as explain the choice of '7' and the index 'Kp'.

We also require the knowledge that to measure the state of the atmosphere at a particular altitude that certain instruments, operating in particular modes (e.g. wavelength ranges for optical instruments) are required to sample the thermodynamic and dynamic structure of the neutral atmosphere. A simplified version of this inference is as follows: NeutralAtmosphere is an AtmosphereLayer, which has the property hasState with value restriction PhysicalDomainState, which can initially be all possible parameters. This is combined with one-of restrictions and other inherited local value restrictions to infer a much smaller set of parameters. In the example use case, the choice of neutral atmosphere limits the parameter set from about 800 choices to about 30, and the later choice of the data product further refines the parameter set to between 4-8 options. Each of the remaining parameters have the properties hasSpatialDomain and hasTemporalDomain which are used to determine the spatial and temporal coverage. The spatial coverage addresses the use case requirement of the measurement being towards the arctic circle which in turn is inferred based on the location of the observatory. The temporal coverage is inferred from the time of the high geomagnetic activity, discussed above. We encode all asserted relations in OWL and utilize reasoners to make the required inferences.

4 Leveraging Semantic Technologies

The initial prototype VSTO software design (which has undergone one evolution to date) is organized into several clearly defined and separated logical layers.

OWL Ontologies: the set of ontologies describing the major classes of objects and their interrelationships. We used [Protégé] and [SWOOP] to develop and browse the ontologies. For the purposes of distributed and extensible design, we had a modular structure including specific ontologies that described (and extended) particular integrated data services (such as CEDAR and MLSO) as well as core ontologies for use in all of our VO projects.

Object Model: We used the Protégé environment tools to generate a hierarchy of Java classes complete with class stub extensions that may be used to insert custom functionality (for example, for executing specific queries versus a database repository). The VSTOfactory class (also created automatically) is used to create instances of the Java classes, which are the equivalent of the OWL individuals.

Services: VSTO-specific Java service classes were developed to provide a high-level Object-Oriented API to query the VSTO knowledge basis (for example, to retrieve all instances of an Instrument that are operated by a given Observatory).

The current VSTO architecture utilizes the [Jena] and [Eclipse] plug-ins for Protégé to generate the Java stub code for the ontology classes and allows the incorporation of existing calls to the CEDAR catalog service for the date and time coverage for the data from the instruments (the remainder of the previous calls to the catalog, implemented in [mySQL], are encoded as individuals in the ontology). The user interface is built on the [Spring] framework, which encodes the workflow and navigation features. Examples of the prototype implementation are displayed [Fox,

McGuinness et al. 2006]. The initial implementation uses the Pellet [Sirin, et al, 2006] reasoner, which will operate on over 10,000 triples and typically returns results in a few seconds on our deployment platform.

Our implementation utilizes an existing set of services for returning selections over a large number (over 60 million records) of date/time information in the CEDAR database. We also utilize a set of existing services for plotting the returned data, which are currently operating in the production CEDARWEB. These services utilize the Interactive Data Language [IDL] as well as the Open Source Project for Network Data Access Protocol [OPeNDAP] to access the relevant data elements from the data archive. The ability to rapidly re-use these services is an essential and effective tool in our effort to deploy a production data-driven virtual observatory environment.

5 Discussion

One of the overriding requirements for virtual observatories is to be able to find and retrieve a wide variety of data sources. As a result, the ability to rapidly develop the semantic framework, deploy and test it is essential. Fortunately, the availability of the OWL language, and related environments and reasoners supported rapid ontology building along with reasoning and queries for testing. In this section, we will highlight some of the positive and negative aspects of our journey applying semantic technologies in Virtual Observatory Settings.

From a representation and reasoning perspective, the existing OWL-DL language and its associated reasoners essentially met our primary needs. Our main concerns for modeling included encoding interconnected class hierarchies with numerous properties (both data and object) with a rich set of value restrictions. Our main concerns from a reasoning perspective include inheritance of property restrictions, limited disjoint and enumerated class reasoning, and enforcing domain and range statements.

The two representational requirements that we need to work on with time include:

1. more extensive support for numeric representation and comparison, and
2. support for modeling typical / default values.

These representational issues are not negatively impacting our current implementation and deployed systems but we will need to handle more complete descriptions and reasoning requirements over time.

We also need to encode provenance meta information concerning our data. Our primary needs at this point reflect data provenance (in line with [Buneman, et al, 2001] – where the data actually came from) although it is quickly moving to knowledge provenance – where the data came from *AND* how it was manipulated. This is in line with what is captured in the proof markup language (PML [Pinheiro da Silva, et al, 2006]) and what is manipulated and presented in the Inference Web Explanation Architecture [McGuinness and Pinheiro da Silva, 2004]. We introduced the notion and distinctions of knowledge provenance in [Pinheiro da Silva, et al, 2003]. Today, we are not capturing this provenance information nor providing search and filtering based on it, but in time, we expect to require this capacity. The initial design includes capturing provenance related to datasets and the instruments (and

error ranges) included in them. The next level is to include knowledge provenance for the actual deductions.

From an environmental perspective, the tool support for *individual* user development was useful and adequate. We heavily used OWL editors, reasoners, and some plug-ins for Protégé for generating java. We lacked supportive collaborative development and analysis tools.

Over the long run, we will need to develop and maintain broad and deep ontologies The ontology maintenance and evolution will need to be carried out by the community. Initially, we just need better support for small team collaborative ontology evolution efforts. Over time, we will need support for widely distributed contributions to ontology maintenance.

One interesting development in our work, as in many other projects like ours, was that we had no choice but to integrate many controlled vocabularies into our ontology. Our data services, with which we had to integrate, already had made choices about using either recognized or defacto standard vocabularies. We originally thought we would rely more heavily on a large background ontology for earth and space sciences – SWEET. Since our initial efforts have been somewhat well defined and also in areas where our team includes leading experts, there has been less need to do a complete import of the entire SWEET vocabulary. Instead, our effort has been focused on using the same terms and in the same way as SWEET when it fits our effort, but NOT to import the entire background ontology. This is largely because, while the large background ontology is a well respected source, it both does not have nearly enough detail in some areas that are critical to our effort, and simultaneously, it contains way more terminology than our effort requires. Importing a large background ontology can be a dual-edged sword and in our original effort, we have found it better to build critical core components as driven by our use cases, staying informed of other resources but not solely relying on them. This has not been a style of work unique to our effort. The same pattern has played itself out with many efforts that considered, for example, whether they should import large upper ontologies (such as SUO or DOLCE) or large mid level ontologies. We note this issue since we expect that many efforts will make similar tradeoffs and thus the need for search tools (such as SWOOGLE [Finin, et al, 2005]), and merging and analysis tools (such as Chimaera [McGuinness, et. al, 2000]), will grow with time.

6 Summary

We designed and implemented an initial semantic data framework for virtual observatories. We leveraged semantic technologies to help provide semantic integration. Our ontology-enhanced services and tools provide retrieval, analysis, and plotting support. We have deployed our implementations for solar and solar-terrestrial information services for CEDAR and Mauna Loa Solar Observatory [MLSO]. While there have been some challenges to using the new technologies, we have found that semantic technologies provide a technological advantage, especially when trying to function in widely distributed, broad, and evolving data settings. We

believe semantic technologies provide a foundation for the evolving and growing trend of work in scientific data integration and virtual observatories.

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. VSTO - http://vsto.hao.ucar.edu/
2. Peter Fox, Deborah L. McGuinness. Semantically-Enabled Large-Scale Science Data Repositories, ISWC Semantic Web in Use Track, Athens, Ga, LNCS, in press, Nov., 2006.
3. SWEET - http://sweet.jpl.nasa.gov/
4. CEDAR - http://cedarweb.hao.ucar.edu/instruments/mfp.html
5. Protégé - http://protege.stanford.edu/
6. SWOOP - http://www.mindswap.org/2004/SWOOP/
7. Jena - http://jena.sourceforge.net/
8. Eclipse - http://www.eclipse.org/
9. MySQL - http://www.mysql.org/
10. Spring - http://www.springframework.org/
11. 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*
12. IDL - http://www.ittvis.com/
13. OPeNDAP - http://www.opendap.org/
14. Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and Where: A Characterization of Data Provenance. In Proceedings of 8th Intl. Converence on Database Theory, pp 316-330, January, 2001.
15. 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.
16. 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.
17. 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
18. Paulo Pinheiro da Silva, Deborah L. McGuinness and Richard Fikes. A Proof Markup Language for Semantic Web Services. Information Systems, Volume 31, Issues 4-5, June-July 2006, Pages 381-395.
19. Paulo Pinheiro, Deborah L. McGuinness, Rob McCool. Knowledge Provenance Infrastructure. In Data Engineering Bulletin Vol 26, No. 4, pages 26-32, December 2003. http://www.ksl.stanford.edu/people/dlm/papers/provenanceinfrastructure.pdf
20. MLSO - http://mlso.hao.ucar.edu/