The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research

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

[1] Knowledge Systems, Artificial Intelligence Laboratory, Stanford University, 345 Serra Mall, Stanford, CA 94305, {dlm@cs.stanford.edu}

[2] McGuinness Associates, 20 Peter Coutts Circle, Stanford, CA 94305, {dlm, jbenedict}@mcguinnessassociates.com

[3] High Altitude Observatory, Earth Sun Systems Lab, National Center for Atmospheric Research[#], PO Box 3000, Boulder, CO 80307, {pfox,pwest,jgarcia}@ucar.edu

[4] Scientific Computing Division, Computing and Information Systems Lab, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, {luca,don}@ucar.edu

[#] The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research with substantial sponsorship from the National Science Foundation.

Abstract

The Virtual Solar-Terrestrial Observatory is a production semantic web data framework providing access to observational datasets from fields spanning upper atmospheric terrestrial physics to solar physics. The observatory allows virtual access to a highly distributed and heterogeneous set of data that appears as if all resources are organized, stored and retrieved/used in a common way. The end-user community comprises scientists, students, data providers numbering over 600 out of an estimated community of 800. We present details on the case study, our technological approach including the semantic web languages, tools and infrastructure deployed, benefits of AI technology to the application, and our present evaluation after the initial nine months of use.

1. Introduction

Scientific data is being collected and maintained in digital form in high volumes by many research groups. The need for access to and interoperability between these repositories is growing both for research groups to access their own data collections but also for researchers to access and utilize other research groups' data repositories in a single discipline or, more interestingly, in multiple disciplines. The promise of the true virtual interconnected heterogeneous distributed international data repository is starting to be realized. But there exist many challenges including interoperability and integration between data collections. We are exploring ways of technologically enabling scientific virtual observatories - distributed resources that may contain vast amounts of scientific observational data, theoretical models, and analysis programs and results from a broad range of disciplines. Our goal is to make these repositories appear as if they are one integrated local resource, while realizing that the information is collected by many research groups, using a multitude of instruments with varying instrument settings in multiple experiments with different goals, and captured in a wide range of formats. Our setting is interdisciplinary virtual observatories. By definition, a researcher with a single Ph.D. is unlikely to have enough depth to be considered a subject matter expert in the entire collection. Vocabulary differences across disciplines; varying terminologies, similar terms with different meanings, and multiple terms for the same phenomenon or process provide challenges. These challenges present barriers to efforts that hope to use existing technology in support of interdisciplinary data query and access. They present even greater barriers when the interdisciplinary applications must go beyond search and access to actual manipulation and use of the data. We used artificial intelligence technologies, in particular semantic technologies, to create declarative, machine operational encodings of the semantics of the data to facilitate interoperability and semantic integration of data. We then semantically enabled web services to find, manipulate, and present scientific data.

Encoding formal semantics in the technical architecture of virtual observatories and their associated data frameworks is similar to efforts to add semantics to the web in general [Berners-Lee et al. 2006], workflow systems [Gil et al. 2006], computational grids [DeRoure et al. 2005] and data mining frameworks [Rushing et al. 2005]. The value added by basic knowledge representation and reasoning is supporting both computer to computer and researcher-to-computer interfaces that find, access and use data in a more effective, robust and reliable way.

We describe our Virtual Observatory project, including our vision, design and AI-enabled implementation. We will highlight where we are using Semantic Web technologies and discuss our motivation for using them

and some benefits we are realizing. We describe our deployment and maintenance settings that started production in the summer of 2006.

2. Task Description

Our goal was to create a scalable interdisciplinary Virtual Solar-Terrestrial Observatory [VSTO] for searching, integrating, and analyzing distributed heterogeneous data resources. A distributed multi-disciplinary internet-enabled virtual observatory requires a higher level of semantic interoperability than was previously required by most (if not all) distributed data systems or discipline-specific virtual observatories. Existing work targeted subject matter experts as end users and did little to support integration of multiple collections (other than providing basic access to search interfaces that are typically specialized and idiosyncratic).

Our science domains utilize a balance of observational data, theoretical models, analysis, and interpretation to make effective progress. Since many data collections are interdisciplinary, and growing in volume and complexity, the task of making them a research resource that is easy to find, access, compare and utilize is still a significant challenge. These collections 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. We view VSTO as representative of multi-disciplinary virtual observatories in general and thus claim that many of our results can be applied in other multi-disciplinary VO efforts.

In order to provide a scientific infrastructure that is usable and extensible, VSTO requires contributions concerning semantic integration, and knowledge representation while requiring depth in a number of science areas. We chose an AI technology foundation because of the promise for a declarative, extensible, re-usable technology platform.

3. Application Description

The application uses background information about the terms used in the subject matter repositories. We encoded this information in OWL [McGuinness & van Harmelan, 2004]. We used both the SWOOP[1] and Protégé[2] editors for ontology development. The definitions in the ontologies are used (via the Jena[3] and Eclipse[4] Protégé plug-ins) to generate java classes in a java object model. We built java services that use this java code to access the catalog data services. We use the PELLET[5] reasoning engine to compute information that is implied and also to identify contradictions. The user interface uses the Spring[6] framework for supporting workflow and navigation.

The main AI elements that support the semantic foundation for integration in our application include the OWL ontologies and a description logic reasoner (along with supporting tool infrastructure for ontology editing and validation). We will describe these elements, how they are used to create "smart" web services, and their impact in the next two sections.

4. Artificial Intelligence Technology Usage Highlights

4.1 Ontologies

We made the effort to create ontologies defining the terms used in the data collections because we wanted to leverage the precise formal definitions of the terms for semantic search and interoperability. The use cases described below were used to scope the ontologies. The general form of the use cases is "retrieve data (from appropriate collections) subject to (stated and implicit) constraints and plot in a manner appropriate for the data. The three initial motivating use case scenarios are provided below in a templated form and then in an instantiated form:

Template 1: Plot the values of parameter X as taken by instrument Y subject to constraint Z during the period W in style S.

Example 1: Plot the Neutral Temperature (Parameter) taken by the Millstone Hill Fabry-Perot interferometer (Instrument) looking in the vertical direction from January 2000 to August 2000 as a time series.

Template 2: Find and retrieve image data of the type X 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, representing the state of the neutral atmosphere anywhere above 100km and toward the Arctic circle (above 45N) at times of high geomagnetic activity.

After we elaborated upon the use cases, we identified the breadth and depth of the science terms that were used to determine what material we needed to cover and also to scope the search for controlled vocabulary starting points. Essentially we looked at the variables in the templates above and natural hierarchies in those areas (such as an instrument hierarchy), and important properties (such as instrument settings), and restrictions. We also looked for

[1] http://www.mindswap.org/2004/SWOOP/
[2] http://protege.stanford.edu/
[3] http://jena.sourceforge.net/
[4] http://www.eclipse.org/
[5] http://www.mindswap.org/2003/pellet/
[6] http://www.springframework.org/

useful simplifications in areas, such as temporal domain. The data collections already embodied a significant number of controlled vocabularies. The CEDAR[1] program – one of our motivating scientific communities – embodies a controlled vocabulary including terms related to observatories, instruments, operating modes, parameters, observations, etc. Another motivating scientific community – the Mauna Loa Solar Observatory solar atmospheric physics observations – also embodies a controlled vocabulary with significant overlap.

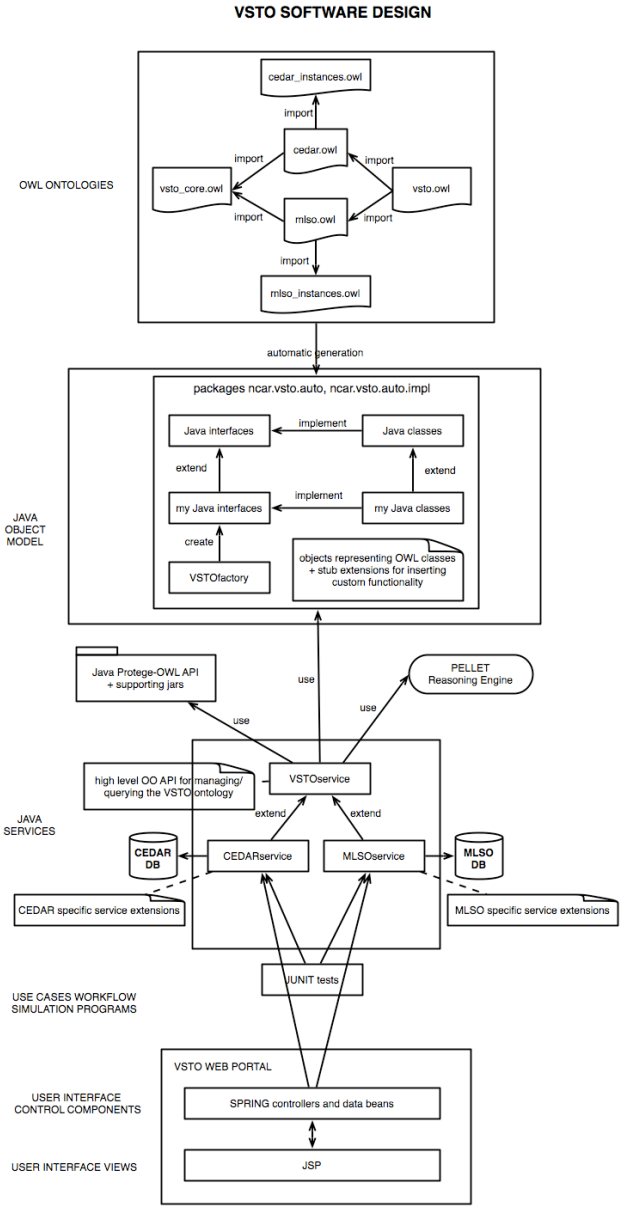

Figure 1: VSTO software architecture.

We searched for existing ontologies in our domain areas and identified SWEET an ontology gaining traction in the science community with significant overlap. This ontology covered *much* more than we needed in breadth, and not enough in depth in multiple places. We reused the conceptual decomposition and terms from the ontology as much as possible and added depth in the areas we required.

We focused on high leverage domain areas. Our first focus area was instruments. One challenge for integration of scientific data taken from multiple instruments is understanding the data collection conditions. 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). A more sophisticated notion is capturing the assumptions embedded in the experiment in which the data was collected and potentially the goal of the experiment. Phase II of our work will address these latter issues. A schematic of part of the ontology is given in Figure 2.

4.2 Reasoning

Our goal was to create a system usable by a broad range of people, some of whom will not be trained in all areas of science covered in the collection. The previous systems required a significant amount of domain knowledge to formulate meaningful and correct queries. Previous interfaces required multiple decisions (8 for CEDAR and 5 for MLSO) to be made by the query generator and those decisions were difficult to make without depth in the subject matter. We used the background ontologies together with the reasoning system to do more work for users and to help them form queries that are both syntactically correct and semantically meaningful. For example, in one work flow pattern, a user is prompted for an instrument and they may choose to filter the instruments by class. If, they ask for photometers, they will be given options shown in Figure 3, at least some of which, would not be obvious by name that they can act as a photometer.

An unexpected outcome of the additional knowledge representation and reasoning was that the same data query workflow is used across the two disciplines. We expect it to generalize to a variety of other datasets as well and we have seen evidence supporting this expectation in our work on other semantically-enabled data integration efforts in domains including volcanology, plate tectonics, and climate change [Fox et al. 2006b].

The reasoner is also used to deduce the potential plot type and return products as well as the independent variable for plotting on the axes.. Previously, users needed to specify all of these items without assistance. One useful reasoning calculation is the determination of parameters that make sense to plot along with the parameter specified. The background ontology is leveraged to determine, for example, that if one is retrieving data concerning neutral temperature (subject to certain conditions) that a time series plot is the appropriate

[1] http://www.cedarweb.hao.ucar.edu

plotting method and neutral winds (the velocity field components) should be shown.

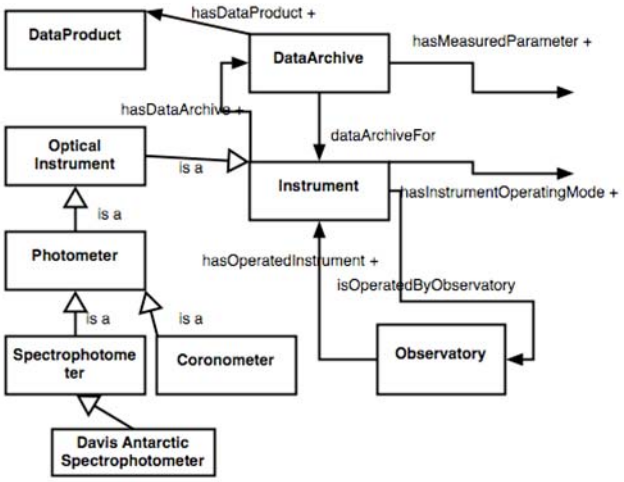

Figure 2: VSTO ontology instrument fragment

4.3 Complex Scientific Data Case Study

Our first and third use cases involve a heterogeneous collection of community data from a nationally funded global change research program - CEDAR. The data collection comprises over 310 different instruments, and the data holdings, which are often specific to each instrument, contains over 820 measured quantities (or parameters) including physical quantities, derived quantities, indices, and ancillary information. CEDAR is further complicated by the lack of specification of independent variables in datasets. Also, the original logical data record encoding for many instruments contains interleaved records representing data from the instrument operating in different modes. Thus odd and even records typically contain different parameters. Sometimes these records are returned without column headings so the user needs to be knowledgeable in the science domain and in the retrieval system just to make sense of the data.

In solar physics images, the original data presentation was that of complex data products, e.g. Mark IV White Light Polarization Brightness Vignetted Data (Rectangular Coordinates). This is a compound description containing Instrument name (Mark IV), parameter (Brightness), operating mode (White Light Polarization), and processing operations (Vignetted Data indicates it has not been corrected for that effect, and a coordinate transformation to rectangular coordinates). Further, the data content retrieved cannot be distinguished from another file unless the filename encoding is understood.

4.4 Ontologies for Interdisciplinary Observational Science Systems

We focused on six root classes: Instrument, Observatory, Operating Mode, Parameter, Coordinate (including Date/Time and Spatial Extent) and Data Archive. While this set of classes does not cover all observational data, it was interesting to note that as we added data sources to the VSTO use cases, we have found these classes to capture the key and defining characteristics of a significant number of observational data holdings in solar and solar-terrestrial physics. As a result, the knowledge represented in these classes is applicable across a range of disciplines. While we do not claim that we have designed a universal broad coverage representation for all observational data sources we believe that this is a major step in that direction and has strong similarities to work in the geo-spatial application domain [Cox 2006, Wolff et al. 2006).

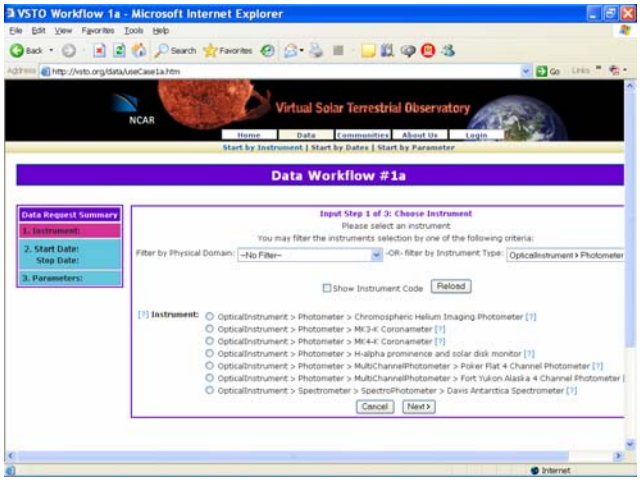

Figure 3: VSTO data search and query interface, exposing taxonomy-based instrument selection.

5 Uses of AI Technology: Ontology-Enhanced Search

VSTO depends on background ontologies, reasoners, and from a maintenance perspective, the supporting semantic technology tools including ontology editors, validators, and plug-ins for code development**.** We designed the ontology to use only the expressive power of OWL-DL rather than moving to OWL-Full so that we could leverage the reasoners available for OWL-DL. Within OWL-DL, we basically had the expressive power we needed with the following two exceptions. We could use support for numerics (representation and comparison) and defaults.. The current application does not use an encoding for default values. Our current application handles numerical analysis with special purpose query and comparison code. While it would have been nice to have more support within the semantic web technology toolkit, this issue is some what less of an issue for our application since the sheer quantity of numerical data meant that we needed special purpose handling anyway. The quantity of date data in the distributed repositories is overwhelming, so we have

support functions for accessing it directly from those repositories instead of actually retrieving it into some cached or local store. Our solution uses semantically-enhanced web services to retrieve the data directly.

We used only open source free software for our project. From an ontology editing and reasoning perspective, this mostly met our needs. A few times in the project, it would have been nice to have the support that one typically gets with commercial software but we did get some support where needed on the mailing lists and with limited personal communication. The one thing that we would make the most use of if it existed would be a commercial strength collaborative ontology evolution and source control system. Our initial rounds of development on the ontology were distributed in design but centralized in input because our initial environment was fragile in terms of building the ontology and then generating robust functional java code. The issues concerning the development environment did eventually get resolved and we are now doing distributed ontology development and maintenance using modularization and social conventions.

6 Application Use and Evaluation

VSTO has been operational since the summer of 2006. It has achieved broad acceptance and is currently used by approximately 80% of the research community[1]. The production VSTO portal has been the primary entry point to date for users (as well as those interested in semantic web technologies in practice). Until recently, all data query formations up to the stage of data retrieval in the new and old portal were treated anonymously. The newest release of the portal now captures session statistics which we will report at the meeting in July. We now collect query logs in the form of both accesses to the triple store (Jena in memory), as well as calls to the reasoner (Pellet) and any SPARQL queries. We are also investigating click-stream methods of instrumenting parts of the portal interface as well as the underlying key classes in the API. Our intent is to capture and distinguish between portal and web services access (which also record details of the arguments and return documents) and query formation.

Currently there are on average between 80-90 distinct users authenticated via the portal and issuing 400-450 data requests per day, resulting in data access volumes of 100KB to 210MB per request. In the last year, 100 new users have registered, more than four times the number from the previous year. The users registered last year when the new portal was released, and after the primary community workshop in June at which the new VSTO system was presented. At that meeting, community agreement was given to transfer operations to the new system and move away from the existing one.

At the 2006 CEDAR workshop a priority-area for the community was identified which involved the accuracy and consistency of temperature measurements determined from instruments like the Fabry-Perot Interferometer and as a result, we have seen a 44% increase in data requests in that area. We increased the granularity in the related portion of the ontology to facilitate this study. We focused on improving a users' ability to find related, or supportive data, with which to evaluate the neutral temperatures under investigation. We are seeing an increase (10%) in other neutral temperature data accesses, which we believe is a result of this related need.

One measure that we hoped to achieve is to have usage by all levels of domain scientist – from the PI to the early level graduate student. Anecdotal evidence shows this is happening and self classification also confirms the distribution. A scientist doing model/observational comparisons: noted "took me two passes now, I get it right away", "nice to have quarter of the options", and "I am getting closer to 1 query to 1 data retrieval, that's nice".

Additionally, members of our team who do not have training in the subject area are able to use this interface while they were unable to use previously existing systems (largely because they did not have enough depth in the area for example to know which parameters needed to be grouped together or other subject-specific information). As we presented this work in computer, biomedical, and physical science communities, we have had many interested parties request accounts to try out the capabilities and all have successfully retrieved or plotted data, even users from medical informatics who know nothing about space physics. One commented "This is cool, I can now impress my kids". This was made possible by appropriately plotting the data in a visually appealing and meaningful way, something that someone unfamiliar with the data or science could not have done before.

There have been multiple payoffs for the system many of which have quantitative metrics:

1. Decreased input requirements: The previous system required the user to provide 8 pieces of input data to generate a query and our system requires 3. Additionally, the three choices are constrained by value restrictions propagated by the reasoning engine. Thus, we have made the workflow more efficient and reduced errors (note the supportive user comments two paragraphs above)

2. Syntactic query support: The interface generates only syntactically correct queries. The previous interface allowed users to edit the query directly, thus providing multiple opportunities for syntactic errors in the query formation stage. As one user put it: "I used to do one query, get the data and then alter the URL in a way I thought would get me similar data but I rarely succeeded, now I can quickly re-generate the query for new data and always get what I intended".

[1] We determined this percentage by taking the number of people in the community as measured by the most recent subject matter conferences and the number of registered users for our system.

3. Semantic query support: By using background ontologies and a reasoner, our application has the opportunity to only expose query options that will not generate incoherent queries. Additionally, the interface only exposes options for example in date ranges for which data actually exists. This semantic support did not exist in the previous system. In fact we limited functionality in the old interface to minimize the chances of misleading or semantically incorrect query construction. This means for example, that a user has increased functionality – i.e., they can now initiate a query by selecting a class of parameter(s). As the query progresses, the sub-classes and/or specific instances of that parameter class are available as the datasets are identified later in the query process. We removed the parameter initiated search in the previous system because only the parameter instances could be chosen (for example there are 8 different instances that represent neutral temperature, 18 representations of time, etc.) and it was too easy for the wrong one to be chosen, quickly leading to a dead-end query and frustrated user. One user with more than 5 years of CEDAR system experience noted: "Ah, at last, I've always wanted to be able to search this way and the way you've done it makes so much sense".

4. Semantic integration: Users now depend on the ontologies rather than themselves to know the nuances of the terminologies used in varying data collections. Perhaps more importantly, they also can access information about how data was collected including the operating modes of the instruments used. "The fact that plots come along with the data query is really nice, and that when I selected the data it comes with the correct time parameter" (New graduate student, ~ 1 year of use). The nature of the encoding of time for different instruments means that not only are there 18 different parameter representations but those parameters are sometimes recorded in the prologue entries of the data records, sometimes in the header of the data entry (i.e. as metadata) and sometimes as entries in the data tables themselves. Users had to remember (and maintain codes) to account for numerous combinations. The semantic mediation now provides the level of sensible data integration required.

5. Broader range of potential users: VSTO is usable by people who do not have PhD level expertise in all of the domain science areas, thus supporting efforts including interdisciplinary research. The user population consists of: Student (under-graduate, graduate) and non-student (instrument PI, scientists, data managers, professional research associates). For CEDAR, students: 168, non-students: 337, for MLSO, students: 50, non-students: 250. In addition 36% and 25% of the users are non-US based (CEDAR – a 57% increase over the last year - and MLSO respectively). The relative percentage of students has increased by ~10% for both groups.

Over time, as we continue to add data sources and their associated instruments, and measured parameters, users will benefit by being able to find even more data relevant to their inquiry than before with no additional effort or changes in search behavior. For example, both dynamic and climatological models to be added, provide an alternate, complementary or comparative source of data to those measured by instruments but at present a user has to know how to search for and use these data. Our approach to developing the ontology allows us to add new subclasses, properties, and relationships, in a way that will naturally evolve the reasoning capabilities available to a user, as well as to incoming and outgoing web services, especially as those take advantage of our ontologies.

We conducted an informal user study asking three questions: What do you like about the new searching interface? Are you finding the data you need? What is the single biggest difference? Users are already changing the way they search for and access data. Anecdotal evidence indicates that users are starting to think at the science level of queries, rather than at the former syntactic level. For example, instead of telling a student to enter a particular instrument and date/time range and see what they get, they are able to explore physical quantities of interest at relevant epochs where these quantities go to extreme values, such as auroral brightness at a time of high solar activity (which leads to spectacular auroral phenomena).

We have initiated a more complete user study to be conducted at the annual workshop for the primary user community (CEDAR) held yearly in late June. The results of which will be reported on at the July conference and available from the vsto.org web site following the meeting.

7 Application Development and Deployment

VSTO was funded by a three year NSF grant. In the first year, a small, carefully chosen six person team wrote the use cases, built the ontologies, designed the architecture, and implemented an alpha release. We had our first users within the first 8 months with a small ontology providing access to all of the data resources. Over the last 2 years, we expanded the ontology and made the system more robust and increased domain coverage.

Early issues that needed attention in design included determining an appropriate ontology structure and granularity. Our method was to generate iterations initially done by our lead domain scientist and lead knowledge representation expert, vet the design through use case analysis and other subject matter experts, as well as the entire team. We developed minimalist class and property structures capturing all the concepts into classes and sub-class hierarchies, only including associations, and class value restrictions needed to support reasoning required for the use cases. This choice was driven by several factors: (a) keeping a simple representation allowed the scientific domain literate experts to view and vet the ontology easily: (b) complex class and property relations, while clear to a knowledge engineer, take time for a domain expert to comprehend and agree upon. A practical consideration

arose from Protégé with automatic generation of a Java[tm] class interface and factory classes (see Fig.1 and [Fox et al. 2006a] for details). As we assembled the possible user query workflows and used the Pellet reasoning engine, we built dependencies on properties and their values. If we had implemented a large number of properties and needed to change them or, as we added classes and evolved the ontology, placed properties at a different class levels, the existing code would need to be substantially rewritten manually to remove the old dependencies. Our current approach preserves the existing code, automatically generates the new classes and *adds* incrementally to the existing code. This allows rapid development. Deployment cycles and updates to the ontology can be released with no changes in the existing data framework, benefiting developers and users.

We rely on a combination of editors (Protégé and Swoop). We use Protégé for its plug in support for java code generation. Earlier iterations had some glitches with interoperation in a distributed fashion that supported incremental updates but we overcame these issues and the team now uses a distributed, multi-component platform.

8 Maintenance

Academic and industrial work has been done on ontology evolution environments that this project can draw on. In a paper entitled "Industrial Strength Ontology Management" (Das *et al.*, 2001), a list of ontology management requirements is provided that we endorse and include in our evolution plan: 1. Scalability, 2. Availability, 3. Reliability and Performance, 4. Ease of Use by domain literate people, 5. Extensible and Flexible Knowledge Representation, 6.Distributed Multi-User Collaboration, 7. Security Management, 8. Difference and Merging, 9. XML Interfaces, 10. Internationalization, including support for multiple languages, and 11. Versioning. We would also add: Transparency and Provenance.

Our efforts so far have focused on points 1-3, and to a lesser extent 4, 10, and 11. Our new system needed to be at least as robust and useful as the previously available community system . It was imperative that our application had at least adequate performance, high reliability and availability. We considered two aspects of scaling: (a) expanding to include broader and deeper domain knowledge. (b) handling large volumes of data. We designed for performance in terms of raw quantity of data. We do not import all of the information into a local knowledge base when we know that volumes of data are large; instead we use database calls to existing data services. Thus, we do not achieve decreased performance. We address reasoning performance by limiting our representation to OWL-DL.

We built our ontology design to be extensible and over time, we are finding that the design is holding up both to extension within our project and also to reuse in other projects. We have investigated the reuse of our ontologies in our Semantically-Enabled Science Data Integration project that addresses virtual observatory needs in the overlapping areas of climate, volcano, and plate tectonics. We found that while for example seismologists use some instruments that solar terrestrial physicists do not, the basic properties used to describe the instruments, observatories, and observations are quite similar. Routine maintenance and expansion of the ontologies is done by the larger team.

We promote use cased-based design and extensions. When we plan for extensions, we begin with use cases to identify additional vocabulary and inferences that need to be supported. We have also used standard naming conventions and have maintained as much compatibility as possible with terms in existing controlled vocabularies,

Our approach to distributed multi-user collaboration is a combination of social and technical conventions. This is largely due to the state of the art, where there is no single best multi-user ontology evolution environment. We have one person in charge of all VSTO releases and this person maintains a versioned, stable version at all times. We also maintain an evolving, working version. The ontology is modular so that different team members can work on different pieces of the ontology in parallel.

We are just beginning our work on transparency and provenance. Our design leverages the Proof Markup Language [Pinheiro daSilva, et al, 2006] – an Interlingua for representing provenance, justification, and trust information. Our initial provenance plans include capturing content such as where the data came from. Once captured in PML, the Inference Web toolkit [McGuinness et al, 2004] may be used to display information about why an answer was generated, where it came from, and how much the information might be believed and why.

9 Summary and Discussion

.

We introduced our interdisciplinary virtual observatory project – VSTO. We used semantic technologies to quickly design, develop and deploy an integrated, virtual repository of scientific data in the fields of solar and solar-terrestrial physics. Our new VO can be used in ways the previous system was not conveniently able to be used to address emerging science area topics such as the correctness of temperature measurements from Fabry-Perot Interferometers. A few highlights of the knowledge representation that may be of interest follow.

We designed what appears to be an extensible, reusable ontology for solar-terrestrial physics. It is compatible with controlled vocabularies in use in the most widely used relevant data collections. Further, and potentially much more leverageable, is that the structure of the ontology is withstanding reuse in multiple virtual observatory projects. We have reviewed the ontology with respect to needs for the NSF-funded GEON project, the NASA-funded SESDI project, and the NASA-funded SKIF project.

The SWEET ontology suite was simultaneously much too broad and not deep enough in our subject areas. If we could have imported just the portions of SWEET that we needed and expanded from there, it might have been possible to use more directly. We made every effort to use terms from SWEET and to be compatible with the general modeling style. We are working with the SWEET developers to make a general, reusable, modular ontology for earth and space science. Our ontologies are open source and have been delivered to the SWEET community for integration. A web site is available for obtaining status information on this effort: www.planetont.org.

This project has a multitude of challenges. The scope of the ontology is broad enough that it is not possible for any single scientist to have enough depth in the subject matter to provide the raw content. The project thus must be a collaborative effort. Additionally, a small set of experts could be identified to be the main contributors to particular subject areas and an ontology could be created by them. If the ontology effort stops there though, we will not achieve the results we are looking for. We want to have an extensible, evolving, widely reusable ontology. We believe this requires broad community buy in that will include vetting and augmentation by the larger scientific community and ultimately it needs usage from the broad community and multiple publication venues including a new Journal of Earth Science Informatics.

We also believe judicious work on modularization is critical since our biggest barrier to reuse of SWEET was the lack of support for importing modules that were appropriate for our particular subject areas. We believe this effort requires community education on processes for updating and extending a community resource such as a large (potentially complicated) ontology.

Today, our implementation uses fairly limited inference and supports somewhat modest use cases. This was intentional as we were trying to provide an initial implementation that was simple enough to be usable by the broad community with minimum training. Initial usage reports show that it is well received and that users may be amenable to additional inferential support. We plan to redesign the multiple work flow interface and combine it into a much more general and flexible single work flow that is adaptable in its entry points. Additionally, we plan to augment the ontology to capture more detail for example in value restrictions and thus be able to support more sophisticated reasoning. Additionally, the current implementation has limited support for encoding provenance of data. Thus we will use the provenance Interlingua PML-P to capture knowledge provenance so that end users may ask about data lineage.

Acknowledgements

The VSTO project is funded by the National Science Foundation, Office of Cyber Infrastructure under the SEI+II program, grant number 0431153.

References

Berners-Lee, T, Hall, W., Hendler, J, Shadbolt, N, and Weitzner, J. 2006, Enhanced: Creating a Science of the Web, Science, 313 #5788, pp. 769-771, DOI: 10.1126/science.1126902

Cox, S. 2006, Exchanging observations and measurements data: applications of a generic model and encoding, Eos Trans. AGU Fall Meet., Suppl., 87(52) IN53C-01.

De Roure, D. Jennings, N.R. Shadbolt, N.R. 2005, The semantic grid: past, present, and future, Proceedings of the IEEE, 93, Issue: 3, pp. 669-681, DOI: 10.1109/JPROC.2004.842781.

Fox, P., McGuinness, D.L., Middleton, D., Cinquini, L., Darnell, J.A., Garcia, J., West, P., Benedict, J., Solomon, S. 2006a, 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.. 2006b, Semantically-Enabled Scientific Data Integration. Proceedings of the Geoinformatics Conference, Reston, Virginia, May 10-12, 2006.

Gil, Y., Ratnakar, V. and Deelman, E. 2006, Metadata Catalogs with Semantic Representations, International Provenance and Annotation Workshop 2006 (IPAW2006), Chicago, IL, Eds. L. Moreau and I. Foster, LNCS 4145, pp90-100, Springer-Verlag, Berlin.

Martin, D., Burstein, M., McDermott, D., McGuinness, D., McIlraith, S., Paolucci,.M., Sirin, E.,, Srinivasan, N, and Sycara, K. Bringing Semantics to Web Services with OWL-S. World Wide Web Journal, to appear. Also, Stanford KSL Tech Report KSL-06-21.

McGuinness, D. and Pinheiro da Silva, P. 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. Mylopoulous. 1(4). Fall, 2004.

McGuinness, D., and van Harmelen, F.. OWL Web Ontology Language Overview. World Wide Web Consortium (W3C) Recommendation. February 10, 2004. www.w3.org/TR/owl-features/ .

Pinheiro da Silva, P., McGuinness, D., and Fikes, R. A Proof Markup Language for Semantic Web Services. Information Systems, 31(4-5), June-July 2006, pp 381-395. Prev. version, KSL Tech Report KSL-04-01.

Rushing, J., R. Ramachandran, U. Nair, S. Graves, R. Welch, and A. Lin, "ADaM: A Data Mining Toolkit for Scientists and Engineers," Computers & Geosciences, vol. 31, pp. 607-618, 2005.

Wolff, A., Lawrence, B. N., Tandy, J., Millard, K. and Lowe, D. 2006, Feature Types' as an Integration Bridge in the Climate Sciences, Eos Trans. AGU Fall Meet., Suppl., 87(52) Abstract IN53C-02.