AI Magazine Volume 29 Number 1 (2008) (© AAAI)

Enabling Scientific Research Using an Interdisciplinary Virtual Observatory

The Virtual Solar-Terrestrial Observatory Example

Deborah L. McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict, and Don Middleton

■ *Our work is aimed at enabling a new style of virtual, distributed scientific research. We have designed, built, and deployed an interdisciplinary virtual observatory—an online service providing access to what appears to be an integrated collection of scientific data. The Virtual Solar-Terrestrial Observatory (VSTO) is a production semantic web data framework providing access to observational data sets 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 or used in a common way. The end-user community includes scientists, students, and data providers. We will introduce interdisciplinary virtual observatories and their potential impact by describing our experiences with VSTO. We will also highlight some benefits of the embedded semantic web technology and also provide evaluation results after the first year of use.*

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. Research groups need to access their own increasingly diverse data collections. As investigations begin to include results from many different experiments, researchers also need to access and utilize other research groups' data repositories in a single discipline or, more interestingly, in multiple disciplines. Also, it is not simply trained scientists who are interested in accessing scientific data; lay people are becoming interested in looking at trends in scientific data as well, for example, when they become engaged in climate discussions.

The promise of the true virtual interconnected heterogeneous distributed international data repository is starting to be realized. Many challenges still exist 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. Initially our focus is on trained scientists. Our ultimate goal is to provide support for broader usage, including lay people.

Because we believe science increasingly requires interactions across multiple domains, our setting is interdisciplinary virtual observatories. A researcher with a single Ph.D. is expected to have depth in his or her chosen subject area but 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, especially when the interdisciplinary applications must go beyond search and access to actual manipulation and use of the data. In addition the user community has a more diverse level of education and training. 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 wrote web services that used background knowledge to help them 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 (for example, Gil, Ratnaker, and Deelman [2006]), computational grids (for example, DeRoure, Jennings, and Shadbolt [2005]) and data-mining frameworks (such as 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.

In the rest of this article, 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. We also include results from our initial evaluation study.

Task Description

Our goal was to create a scalable interdisciplinary virtual observatory that would support scientists in searching, integrating, and analyzing distributed heterogeneous data resources. A distributed multidisciplinary 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 initial science domains were those of interest to scientists who study the Earth's atmosphere and the Sun. Our initial virtual observatory is thus VSTO—the Virtual Solar-Terrestrial Observatory. Scientists in these areas must 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 data sets 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 multidisciplinary virtual observatories in general and thus claim that many of our results can be applied in other multidisciplinary virtual observatory efforts.

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, reusable technology platform.

Application Description

The application uses background information about the terms used in the subject matter repositories. We encoded this information in OWL (McGuinness and van Harmelan, 2004)—the recommended web ontology language from the World Wide Web Consortium (W3C). We used best-in-class semantic web tools for development. We used both the SWOOP[1] and Protégé[2] editors for ontology development. The definitions in the ontologies are used (through 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] description logic 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 log-

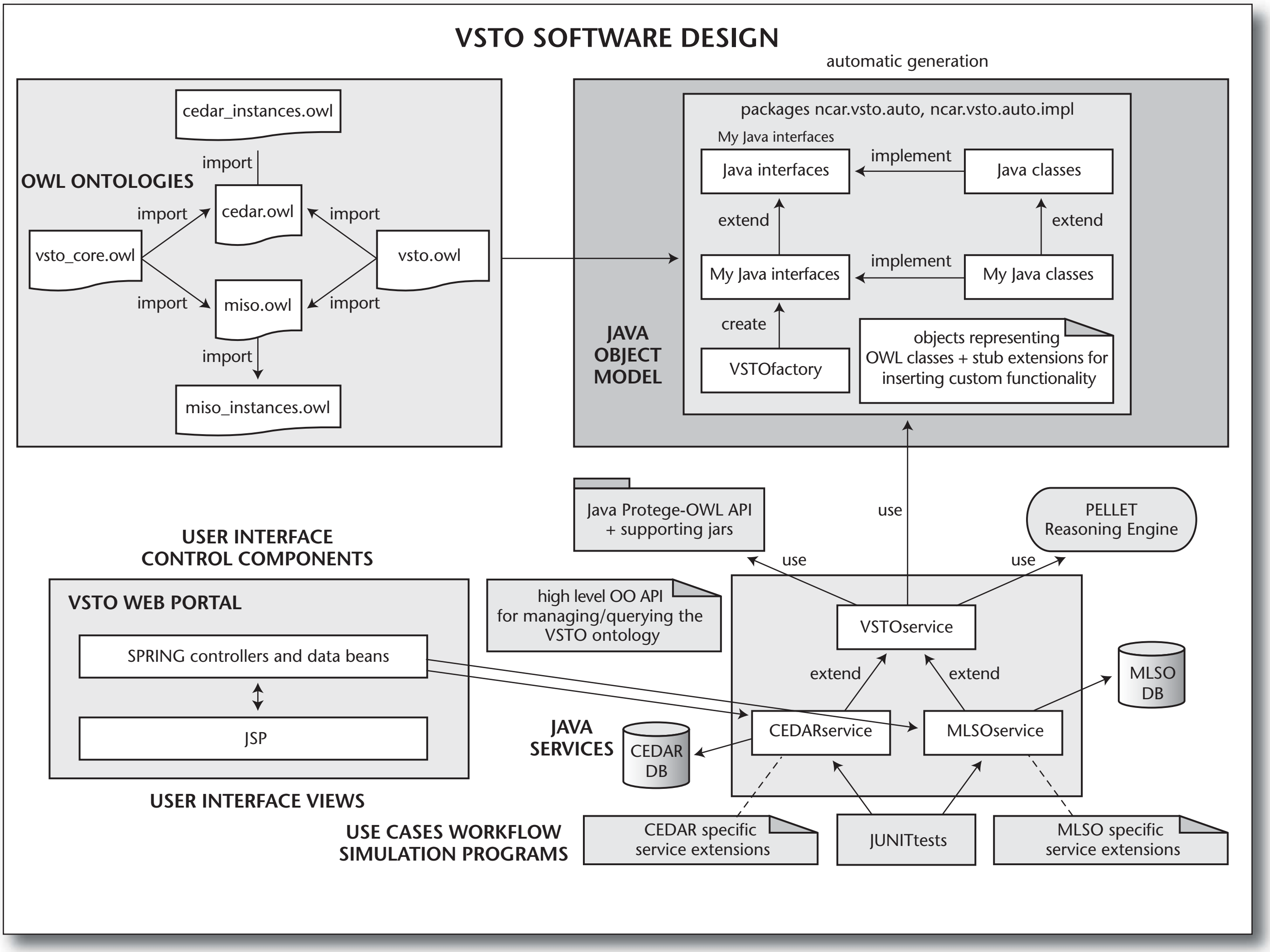


Figure 1: VSTO Software Architecture.

ic 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. Figure 1 depicts the software architecture.

Artificial Intelligence Technology Usage Highlights

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 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* using data product *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 (data product).

Template 2: Find and retrieve image data of 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 100 km and toward the Arctic circle (above 45 degrees N) 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), important properties (such as instrument settings), and restrictions. We also looked for useful simplifications in areas such as the temporal domain. The data collections already embodied a significant number of controlled vocabularies. We began with main science communities: CEDAR[7] and MLSO.[8] The CEDAR archive provides an online database of middle and upper atmospheric, geophysical index, and model data. The MLSO archive provides an online database (including many images) of solar atmospheric physics data. The CEDAR holdings embody a controlled vocabulary including terms related to observatories, instruments, operating modes, parameters, observations, and so on. MLSO holdings also embody a controlled vocabulary with significant overlap in concepts.

We searched for existing ontologies in our domain areas and identified SWEET[9], an ontology gaining traction in the science community with sufficient overlap with our domains. SWEET considers itself as an upper-level ontology for Earth and environmental scientists (and from a general perspective, it would be considered a midlevel ontology). 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. These areas also have proven to be leveragable in applications outside of a solar-terrestrial focus. The expansion into the disciplines of volcanic effects on climate have led us to reuse many of the ontology concepts we developed for VSTO (McGuinness et al. 2007a, Fox et al. 2007a). 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—that is, 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.

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. Initially, we targeted trained scientists (in future work, we plan to expand the interfaces and data collections to include components appropriate for a broader population). The previous science systems required a significant amount of domain knowledge to formulate meaningful and correct queries.

Previous interfaces required multiple decisions (eight for CEDAR and five 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, users are 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, and for at least some of these it is not 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 data sets 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 (McGuinness et al. 2007b, Fox et al. 2006a), which, as noted earlier, is both leveraging our VSTO ontology work and adding the need for additional reasoning in support of the data integration in this latter project (McGuinness et al. 2007a, Fox et al. 2007b).

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) a time series plot is the appropriate plotting method and neutral winds (the velocity field components) should be shown.

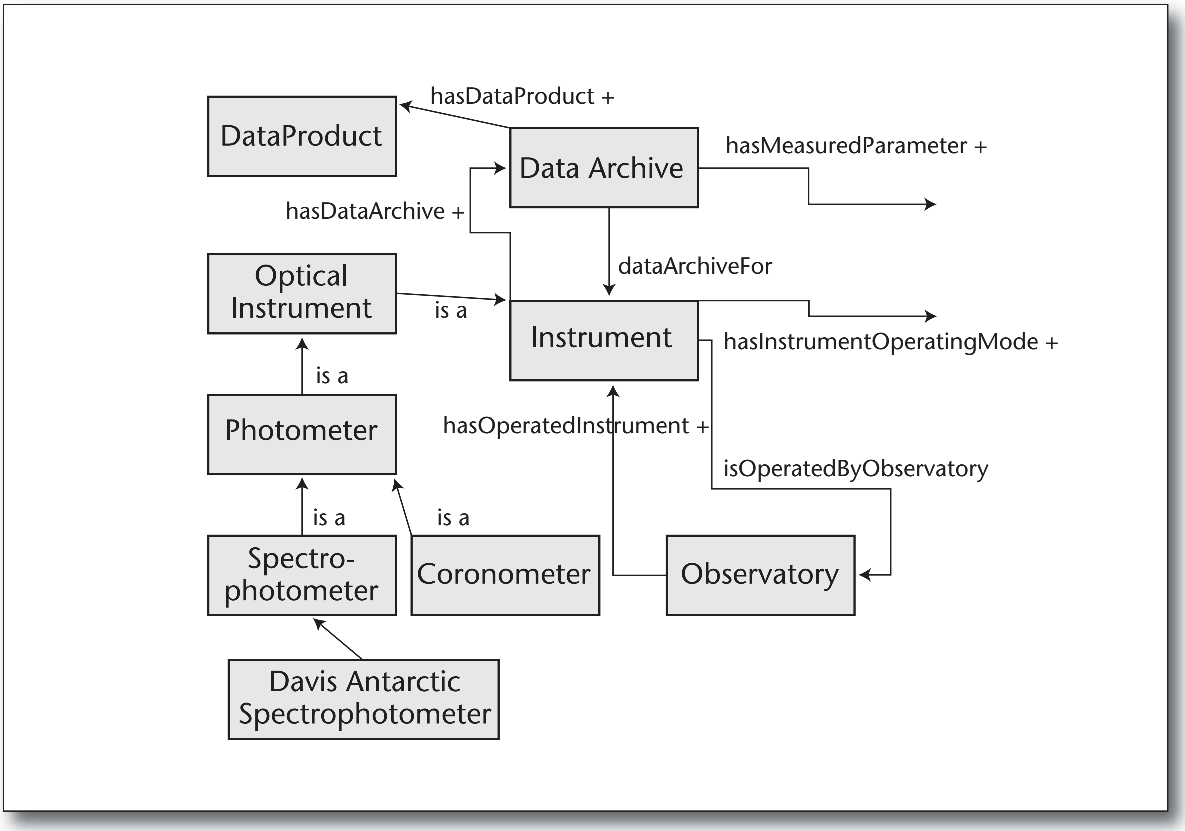


Figure 2: VSTO Ontology Instrument Fragment.

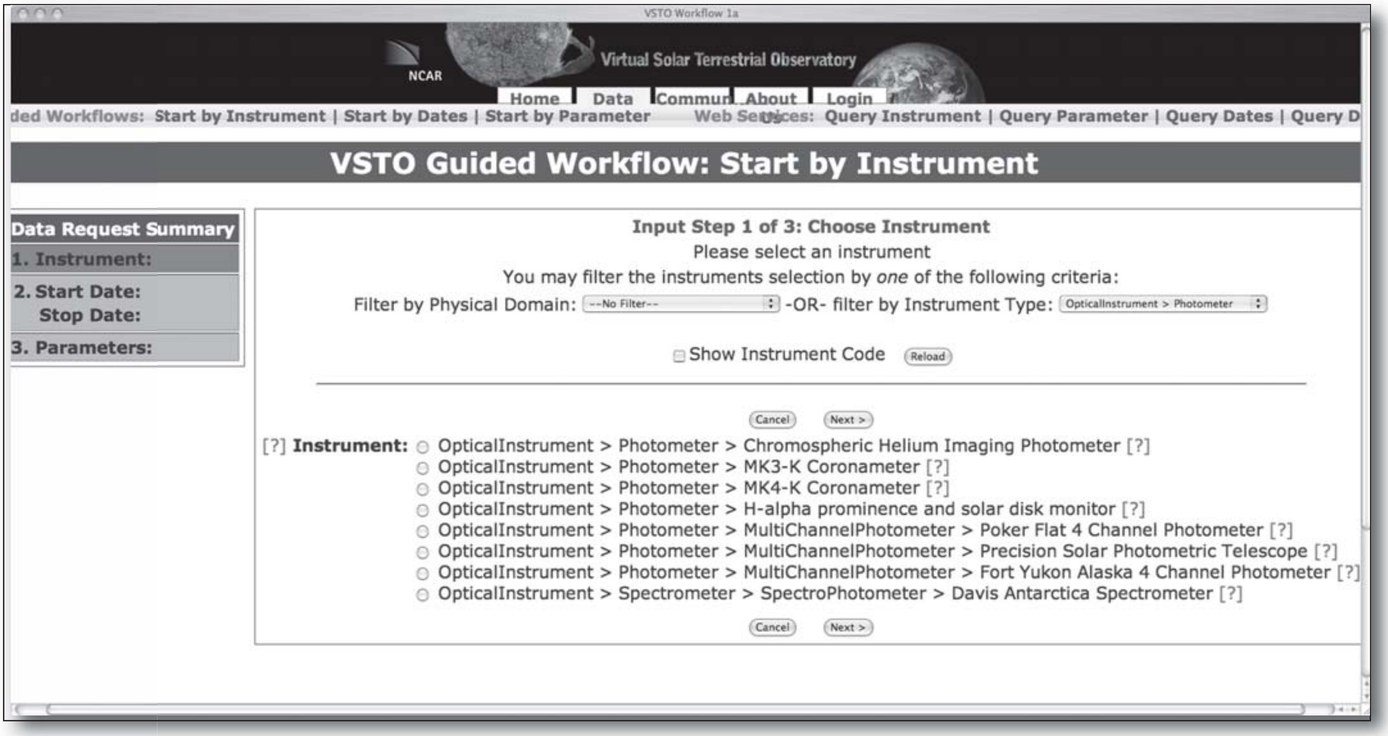
Complex Scientific Data Case Study

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 consists of more than 310 different instruments, and the data holdings, which are often specific to each instrument, contain 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 data sets. 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, for example, 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 is specified). Further, the data content retrieved cannot be distinguished from another file unless the file-name encoding is understood.

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 have 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


VSTO Workflow 1a
NCAR
Virtual Solar Terrestrial Observatory
Home
Data
Commun
About
Login
ded Workflows: Start by Instrument | Start by Dates | Start by Parameter
Web Services: Query Instrument | Query Parameter | Query Dates | Query D
Cancel
Next >
Reload

VSTO Guided Workflow: Start by Instrument

Data Request Summary
1. Instrument:
2. Start Date:
Stop Date:
3. Parameters:

Input Step 1 of 3: Choose Instrument
Please select an instrument
You may filter the instruments selection by *one* of the following criteria:
Filter by Physical Domain: --No Filter-- -OR- filter by Instrument Type: OpticalInstrument > Photometer

□ Show Instrument Code

[?] **Instrument:**
○ OpticalInstrument > Photometer > Chromospheric Helium Imaging Photometer [?]
○ OpticalInstrument > Photometer > MK3-K Coronameter [?]
○ OpticalInstrument > Photometer > MK4-K Coronameter [?]
○ OpticalInstrument > Photometer > H-alpha prominence and solar disk monitor [?]
○ OpticalInstrument > Photometer > MultiChannelPhotometer > Poker Flat 4 Channel Photometer [?]
○ OpticalInstrument > Photometer > MultiChannelPhotometer > Precision Solar Photometric Telescope [?]
○ OpticalInstrument > Photometer > MultiChannelPhotometer > Fort Yukon Alaska 4 Channel Photometer [?]
○ OpticalInstrument > Spectrometer > SpectroPhotometer > Davis Antarctica Spectrometer [?]

Figure 3: VSTO Data Search and Query Interface, Exposing Taxonomy-Based Instrument Selection.

direction. We have tested the hypothesis to some extent in areas including volcanoes and plate tectonics. The work has strong similarities to work in the geospatial application domain (Cox 2006, Wolff et al. 2006).

Uses of AI Technology: Ontology-Enhanced Search

VSTO depends on a number of AI components and tools including background ontologies, reasoners, and—from a maintenance perspective—the semantic technology tools including ontology editors, validators, and plug-ins for code development. We designed the ontology to limit its expressiveness to OWL-DL. We did this so that we could leverage the reasoners available for OWL-DL, along with their better computational efficiency. 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, such as the proposal for numerics in OWL 1.1) 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 is somewhat 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 had 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.

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 percent of the research community.[10] 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 that we reported upon briefly at the IAAI/AAAI June 2007 meeting. 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. Perhaps most importantly the results of an evaluation study conducted at the same meeting are presented in a later section.

Currently there are on average between 80 and 90 distinct users authenticated through the portal and issuing 400 to 450 data requests per day, resulting in data access volumes of 100 KB to 210 MB 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 that involved the accuracy and consistency of temperature measurements determined from instruments like the Fabry-Perot interferometer. As a result, we have seen a 44 percent 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 users' ability to find related or supportive data with which to evaluate the neutral temperatures under investigation. We are seeing an increase (10 percent) 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 principal investigator (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.

First, there are decreased input requirements. The previous system required the user to provide eight pieces of input data to generate a query, and our system requires three. 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 ago).

A second payoff is 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 regenerate the query for new data and always get what I intended."

Third, there is semantic query support. By using background ontologies and a reasoner, our application has the opportunity to expose only query options that will not generate incoherent queries. The interface exposes options, for example, only in date ranges for which data actually exists. This semantic support did not exist in the previous system. In fact, the previous interface had limited functionality to minimize the chances of misleading or semantically incorrect query construction. This means, for example, that users have increased functionality—that is, they can now initiate a query by selecting a class of parameters. As the query progresses, the subclasses and specific instances of that parameter class are available as the data sets 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 eight different instances that represent neutral temperature, 18 representations of time, and so on), 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 five 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.”

Fourth is semantic integration. Users now depend on the ontologies rather than themselves to represent and remember 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, approximately one year of use). The nature of the encoding of time for different instruments means not only that are there 18 different parameter representations but also that those parameters are sometimes recorded in the prologue entries of the data records, sometimes in the header of the data entry (that is, 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.

Finally, there is a broader range of potential users. VSTO is usable by people who do not have Ph.D.-level expertise in all of the domain science areas, thus supporting efforts including interdisciplinary research. The user population consists of students (undergraduate, graduate) and nonstudents (instrument PI, scientists, data managers, professional research associates). For CEDAR, there were 168 students and 337 nonstudents. For MLSO, there were 50 students and 250 nonstudents. In addition 36 percent and 25 percent of the users are non-U.S. based (CEDAR—a 57 percent increase over the last year—and MLSO, respectively). The relative percentage of students has increased by approximately 10 percent 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 the 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 students to enter a particular instrument and date and 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).

A one-hour VSTO workshop was held at the annual CEDAR community meeting on the day after the main plenary presentation for VSTO. The workshop was very well attended with 35 diverse participants (25 were expected) including senior researchers, junior researchers, postdoctoral fellows, and students—including 3 that had just started in the field.

After some self-introductions, eight questions were posed and responses recorded, some by count (yes/no) or comment. Overall responses ranged from 5 to 35 per question. We note some general responses as well as some more specific. Out of these responses we identified some new use cases, which we enumerate. The quantitative questions and answers were:

> How do you like to search for data? Browse, type a query, visual? *Responses:* 10; *Browse* = 7, *Type* = 0, *Visual* = 3.
>
> What other concepts are you interested in using for search, for example, time of high solar activity, campaign, feature, phenomenon, others? *Responses:* 5; all of these, no others were suggested.
>
> Do the interface and its services deliver the functionality, speed, flexibility you require? *Responses:* 30; *Yes* = 30, *No* = 0.
>
> Are you finding the data you need? *Responses:* 35; *Yes* = 34, *No* = 1.
>
> How often do you use the interface in your normal work? *Responses:* 19; *Daily* = 13, *Monthly* = 4, *Longer* = 2.
>
> Are there places where the interface/services fail to perform as desired? *Responses:* 5; *Yes* = 1, *No* = 4.

The more qualitative questions were:

> What do you like about the new searching interface? *Responses:* 9.

What is the single biggest difference? *Responses:* 8.

The general answers were as follows:

1. Fewer clicks to data
2. Autoidentification and retrieval of independent variables
3. Faster
4. Seems to converge faster

The majority of the comments were on the first three, but a few people identified that the search seemed to be more accurate (and converge faster). There were three new users in the session, and their responses were mixed in with the group at large. Interestingly, all three had no problem searching for and accessing the data archives.

Some of the more specific comments (in quote fragment form since they were given orally and the recorder had to write them down quickly) were:

It makes sense now!

[I] Like the plotting.

Finding instruments I never knew about.

Descriptions are very handy.

What else can you add?

How about a python interface [to the services]?

These general and specific comments, also along with the more quantitative answers above, indicate that the VSTO, built on semantic technologies, provided significant additional value for the users and the developers. In several cases, the answers and several unsolicited comments matched almost exactly the ad hoc feedback we had received in the initial study, thus confirming our sense of the initial evaluation.

Several new use cases arose out of the responses. First was the need for a programming/script-level interface, that is, building on the services interfaces, in Python, Perl, C, Ruby, Tcl, and three others. Second was the addition of models alongside observational data, that is, finding data from observations or models that are comparable or compatible. Third were more services (particulary plotting options—for example, coordinate transformation—that are hard to add without detailed knowledge of the data).

The requirement for script-level access was unexpected but afterward seemed a natural way for developers to access the programming-level interface of VSTO (API and web services). Python was the clear first choice but was followed closely by Perl. Numerical (simulation) models are an increasing need for the CEDAR community, especially for integrating models and observational data. The need is that the models fit seamlessly into the selection criteria. This means they are not part of the instrument selection, which is the way the previous interface exposed model simulation data (listing the models as instruments).

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 eight months, with a small ontology providing access to all of the data resources. Over the last two 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 and vet the design through use-case analysis and other subject matter experts, as well as vetting by the entire team. We developed minimalist class and property structures that capture all the concepts into classes and subclass hierarchies including only associations, and class value restrictions needed to support the reasoning required for the use cases. This choice was driven by two factors. First, keeping a simple representation allowed the scientific domain literate experts to view and vet the ontology easily, and second, 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 class interface and factory classes (see figure 1 and Fox et al. [2006a] for details). As we assembled the possible user query work flows 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, had placed properties at different class levels, the existing code would have needed 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, multicomponent platform.

Maintenance

Academic and industrial work has been done on ontology evolution environments that this project can draw on. In a paper titled “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 multiuser 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 on 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 have at least adequate performance, high reliability, and availability. We considered two aspects of scaling, first, expanding to include broader and deeper domain knowledge, and second, 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 (SESDI) project, which 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 are done by the larger team.

We promote use-case-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 multiuser 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 multiuser 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 da Silva, McGuinness, and Fikes 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 and Pinheiro da Silva 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. We have just received National Science Foundation (NSF) funding for extending VSTO in these directions.

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 virtual observatory 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 website 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.

Our follow-up to the initial informal evaluation in a workshop setting provided both general and specific answers and comments, as well as more quantitative yes/no or multiple choices answers. Both sets reaffirmed the sense we obtained in the initial study that our efforts in applying AI techniques in the form of semantics led to an interdisciplinary virtual observatory that provides significant additional value for a spectrum of end users. Perhaps also more importantly it also provides significant additional value for the developers of both the VSTO and other federated virtual observatories and data systems wishing to take advantage of the services that our virtual observatory provides. We plan to engage those developers in articulating the new use cases (for script/programming language access, synthesizing models and observations, and new plotting options) in the near future. We also plan to hold a similar evaluation and feedback workshop at the next annual CEDAR meeting and also the other science communities that VSTO is serving.

Acknowledgements

The VSTO project is funded by the National Science Foundation, Office of Cyber Infrastructure under the SEI+II program, grant number 0431153. This article is an expanded version of a paper presented at the 2007 Innovative Applications of Artificial Intelligence conference.

Notes

1 www.mindswap.org/2004/SWOOP/.

2 protege.stanford.edu/.

3 jena.sourceforge.net/.

4 www.eclipse.org/.

5 www.mindswap.org/2003/pellet/.

6. www.springframework.org/.

7. CEDAR—Coupling, Energetics, and Dynamics of Atmospheric Regions; cedarweb.hao.ucar.edu.

8. MLSO—Mauna Loa Solar Observatory; mlso.hao.ucar.edu.

9. SWEET—Semantic Web for Earth and Environmental Terminologies; sweet.jpl.nasa.gov/ontology/.

10. 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.

References

Berners-Lee, T.; Hall, W.; Hendler, J.; Shadbolt, N.; and Weitzner, J. 2006. Enhanced: Creating a Science of the Web. *Science* 313(5788): 769–771 (DOI 10.1126/Science.1126902).

Cox, S. 2006. Exchanging Observations and Measurements Data: Applications of a Generic Model and Encoding, Eos Transactions of the American Geophysical Union Fall Meeting, Supplement, 87(52) In53c-01.

Das, A.; Wu, W.; and McGuinness, D. L. 2001. Industrial Strength Ontology Management. Stanford Knowledge Systems Laboratory Technical Report KSL-001-09. Stanford University, Stanford CA.

Aseem Das, A.; Wu, Wei; and McGuinness, D. L. 2002. Industrial Strength Ontology Management. In *The Emerging Semantic Web.* ed. I. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, eds. Amsterdam: IOS Press.

De Roure, D.; Jennings, N. R.; and Shadbolt, N. R. 2005. The Semantic Grid: Past, Present, and Future. In *Proceedings of the IEEE,* 93(3): 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.; and Solomon, S. 2006a. Semantically Enabled Large-Scale Science Data Repositories. *The Fifth International Semantic Web Conference (ISWC06), Lecture Notes in Computer Science 4273,* 792–805. Berrlin: Springer-Verlag, Berlin.

Fox, P.; McGuinness, D. L.; Raskin, R.; and Sinha, A. K. 2006b. Semantically Enabled Scientific Data Integration. In *Proceedings of the Geoinformatics Conference*. Washington, DC: American Geophysical Union.

Gil, Y.; Ratnakar, V.; and Deelman, E. 2006. Metadata Catalogs with Semantic Representations. *International Provenance and Annotation Workshop 2006 (Ipaw2006), Lecture Notes in Computer Science* 4145, ed. L. Moreau and I. Foster, 90–100. Berlin: Springer-Verlag.

McGuinness, D.; Fox, P.; Cinquini, C.; West, P.; Garcia, J.; Benedict, J.; and Middleton, D. 2007a, The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In *Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07)*. Menlo Park, CA: AAAI Press.

McGuinness, D.; Fox, P.; Sinha, A. K.; and Raskin, R. 2007b. Semantic Integration of Heterogeneous Volcanic and Atmospheric Data. In *Proceedings of the Geoinformatics 2007 Conference*. Washington, DC: Geological Society of America.

McGuinness, D., and Pinheiro da Silva, P. 2004. Explaining Answers from the Semantic Web: The Inference Web Approach. Special Issue on the International Semantic Web Conference 2003—edited by K. Sycara and J. Mylopoulous. *Web Semantics: Science, Services and Agents on the World Wide Web* 1(4), Fall.

McGuinness, D., and van Harmelen, F. 2004. Owl Web Ontology Language Overview. World Wide Web Consortium (W3c) Recommendation. February 10, 2004 (www.w3.org/TR/owl-features). Cambridge, MA: Massachusetts Institute of Technology.

Pinheiro da Silva, P.; McGuinness, D.; and Fikes, R. 2006. A Proof Markup Language for Semantic Web Services. *Information Systems* 31(4–5): 381–395.

Rushing, J.; Ramachandran, R.; Nair, U.; Graves, S.; Welch, R.; and Lin, A. 2005. Adam: A Data Mining Toolkit for Scientists and Engineers. *Computers and Geosciences* 31(5): 607–618.

Wolff, A.; Lawrence, B. N.; Tandy, J.; Millard, K.; and Lowe, D. 2006. Feature Types as an Integration Bridge in the Climate Sciences. In *Eos Transactions,* American Geophysical Union, Fall Meeting, (San Diego, CA) Supplement, 87(52) Abstract In53c-02. Washington, DC: American Geophysical Union.

Deborah McGuinness is the acting director and senior research scientist of the Knowledge Systems, Artificial Intelligence Laboratory (KSL) at Stanford University and CEO of McGuinness Associates Consulting. By publication time, McGuinness will be the Tetherless World Chair and professor of computer science at Rensselaer Polytechnic Institute (RPI). McGuinness's research focuses on the semantic web, ontologies and ontology evolution environments, knowledge representation and reasoning, explanation, trust, privacy, and search. McGuinness's consulting focuses on helping companies utilize AI research in applications such as smart search, ontology management, information integration, e-science, online commerce, and configuration.

Peter Fox is the chief computational scientist at the High Altitude Observatory, National Center for Atmospheric Research. Fox is a research scientist specializing in the fields of solar-terrestrial physics, computational and computer science, information technology, and grid-enabled, distributed semantic data frameworks. He utilizes state-of-the-art modeling techniques and Internet-based technologies, including the semantic web, and applies them to large-scale distributed scientific data systems. Fox is currently PI for the Virtual Solar-Terrestrial Observatory, the semantically enabled scientific data integration, the semantic provenance capture in data ingest systems, and the CEDAR database projects.

Luca Cinquini holds a Ph.D. in high energy physics from the University of Colorado and is now working as a senior software engineer at the National Center for Atmospheric Research (Boulder, CO). His work focuses on researching and applying web, semantic, and grid technologies to facilitate access to geophysical and space data. He is actively involved in several geoinformatics projects including (among others) the Virtual Solar-Terrestrial Observatory, the Earth System Grid, and the Community Data Portal. He can be reached at luca@ucar.edu.

Patrick West has a bachelor of science in computer science from Indiana University, Bloomington, with 17 years of experience in developing data access and server-side data processing systems. West is a software engineer at the High Altitude Observatory at the University Corporation for Atmospheric Research (HAO/UCAR), developing high-performance server frameworks for the access and processing of scientific data and ontology management.

Jose Garcia has a master of science in computer science and a master of science in applied mathematics from the University of Colorado, with 12 years of experience developing software for scientific and nonscientific projects. Garcia currently works as a software engineer at the High Altitude Observatory developing applications for database systems, ontology management, informatics, numerical analysis, signal processing, high-performance computing, web-based portals, and system integration. His e-mail is jgarcia@ucar.edu.

James Benedict is the chief operating officer of McGuinness Associates Consulting. Benedict has a master's of international management with more than 20 years of work experience in various industries providing management of corporate information technology, accounting systems, and business planning. Benedict helps clients plan, evaluate business impact, develop, deploy, and maintain applications of artificial intelligence and semantic web technology.

Don Middleton leads the Visualization and Enabling Technologies program at the U.S. National Center for Atmospheric Research. He is responsible for a program that encompasses data and knowledge management and systems, advanced analysis and visualization, collaborative visual computing environments, grid computing, and education research and outreach activities. Middleton is currently contributing to the leadership of a number of projects, including the Earth System Grid, the Earth System Curator, the Virtual Solar-Terrestrial Observatory, the North American Regional Climate Change Assessment Project, and the Cyberinfrastructure Strategic Initiative. Middleton has also served on a National Research Council committee, and is currently serving on World Meteorological committees working to build global federated data systems.