Semantic Web Services for Interdisciplinary Scientific Data Query and Retrieval

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

[1] High Altitude Observatory, Earth Sun Systems Lab, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, `{pfox,pwest,jgarcia}@ucar.edu`
[2] Scientific Computing Division, Computing and Information Systems Lab, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, `{luca,don}@ucar.edu`
[3] Knowledge Systems, Artificial Intelligence Laboratory, Stanford University, 345 Serra Mall, Stanford, CA 94305, `{dlm@cs.stanford.edu}`
[4] McGuinness Associates, 20 Peter Coutts Circle, Stanford, CA 94305, `{dlm, jbenedict}@mcguinnessassociates.com`

Abstract

We present a set of four web services provided as a result of our work in developing a semantic data framework in the setting of virtual observatories. These web services allow a client service to search for data using three primary selections: choose parameter, choose date-time range and choose instrument, and also to return appropriate service links to the actual data (the fourth service). These services are built using a shared and common understanding of the inputs, outputs and preconditions as defined by a formal ontology, encoded in OWL-DL and running in an internet accessible environment with Web Service Description Language (WSDL) bindings. Upon invocation the service can utilize reasoning services just as a user of the web portal is able to. The service client can optionally utilize the ontology when it consumes the service for additional knowledge or may be used purely syntactically (as most existing web services are now). We present these services, and how they are developed within a specific domain context for the Virtual Solar-Terrestrial Observatory.

1. *e*-Science

Scientific digital data is being generated, collected and maintained at an ever-increasing rate. To facilitate the carrying out of true *e*-Science there is a need for access to and interoperability amongst the heterogeneous and distributed repositories both within and across traditional community and/or discipline boundaries. Further attributes of *e*-Science cyberinfrastructure include: scalability, sustainability, provenance, seamless data access and integration of virtualized data resources.

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.

A Virtual Observatory (VO) is a suite of software applications on a set of computers that allows users to uniformly find, access, and use resources (data, software, document, and image products and services using these) from a collection of distributed product repositories and service providers. A VO is a service that unites services and/or multiple repositories. [Bentley et al. 2005].

In plain language the main aim of a VO is to make all resources appear to be local and appear to be integrated. This is challenging because 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. In an internet environment, this means providing a means for an incoming user to discover, locate, retrieve and use heterogeneous and perhaps diverse or interdisciplinary data of interest. It also means providing interfaces for incoming requests from user applications and machine generated requests for services.

There are two notable and successful examples of enabling *e*-science using a service-oriented architecture over the internet and in particular using web services. The International Virtual Observatory Alliance (IVOA[1]) with buy-in from large VO efforts such as the National Virtual Observatory (NVO[2]) and ASTROGID[3] and many others, worked to develop a set of 'simple access protocols' to enable application interoperability. Of note are the Simple Image Access Protocol, the Simple Spectrum Access Protocol, the Simple Time Access Protocol and VOTable and VOEvent. Together these web service and XML/SOAP formatted responses are widely used in the astronomy community. The second example is the Open Geospatial Consortium (OGC) who has via their standards process developed protocol standards such as the Web Coverage Service (WCS), the Web Feature Service (WFS) and the Web Map Service (WMS) – together known as WxS. These services are also in wide use in

[1] http://www.ivoa.net/
[2] http://www.us-vo.org/
[3] http://www.astrogrid.org/

applications needing integration via the geospatial coordinate system. We note these two examples in this context since they provide *light-weigh*t semantics via web services. This means that while not containing formal semantic encodings, their names and basis in either particular data-types or coordinate referencing and data product types provide hard coded semantic meaning to those clients accessing them. In essence, terms like image, spectrum, time, coverage, feature and map have a well-defined (and agreed) enough meaning in those communities to provide great utility.

Here we present the web service portion of our work enabling formally-encoded semantic integration of scientific data [Fox, et. al., 2006] within the context of our Virtual Solar-Terrestrial Observatory (VSTO) semantic web application [McGuinness et al. 2007]. VSTO presently covers the fields of solar atmospheric physics and terrestrial middle and upper atmospheric physics. We used semantic web 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 over the internet.

We describe our implementation of the web service as part of our Virtual Observatory project, including the development and use of ontologies to support semantic web services.

2. Use Cases

The use cases described below were used to scope and create the ontologies: to leverage the precise formal definitions of the terms in for semantic search and interoperability. The general forms of the use cases is query for and retrieve data (from appropriate collections) subject to (stated and implicit) constraints and return and/or plot in a manner appropriate for the data. We have 6 use cases but present two motivating use case scenarios relevant to the subject of this paper 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: Expose semantically-enabled, smart data query services via a web services interface allowing composite query formation in arbitrary workflow order.

Example 2: Provide query services for the Virtual Ionosphere-Thermospere-Mesosphere Observatory that retrieve data, filtered by constraints on Instrument, Date-Time, and Parameter in any order and with constraints included in any combination.

Use case 2 is the direct motivation for this paper, but of course in order to achieve use case 2, the web services need to enable use case 1.

3. VSTO as a Semantic Web Application

We were led to encode formal semantics and implement them within the technical architecture of virtual observatories for the same reasons and in the same way that other efforts to add semantics to workflow systems [Gil et al. 2006], computational grids [DeRoure et al. 2005] and data mining frameworks [Rushing et al. 2005]. More detail on the development and the deployment for the VSTO can be found in [Fox et al. 2006, McGuinness et al. 2007]. Initially our focus was on enabling a broad range of users (discipline, experience and science knowledge) via a web portal presence that provides the researcher-to-computer interfaces that find, access and use data of interest. We found multiple payoffs [McGuinness et al. 2007]by utilizing semantic web technologies - decreased input requirements for query: in one base reducing the number of selections from eight to three; the interface generates only syntactically correct queries: which was not always ensurable in previous implementations without semantics; semantic query support: by using background ontologies and a reasoner, our application has the opportunity to only expose coherent query; semantic integration: in the past users had to remember (and maintain codes) to account for numerous different ways to combine and plot the data whereas now semantic mediation provides the level of sensible data integration required; and finally but not least - a broader range of potential users (PhD scientists, students, professional research associates and those from outside the fields).

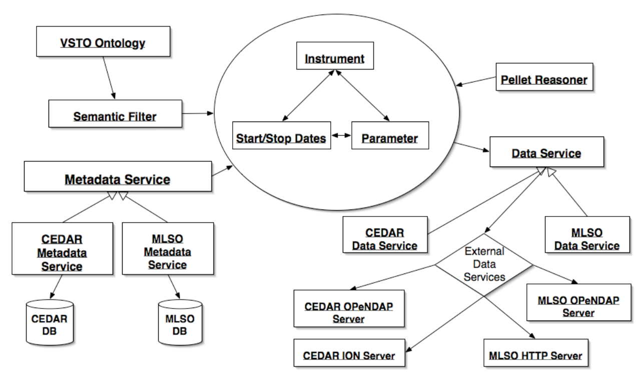

Figure 1: VSTO software / query and access flow architecture.

The main elements that support the semantic foundation for integration in our application include the ontologies and the reasoner along with the supporting tools. We encoded our ontologies in OWL (Ontology Web Language [McGuinness and van Harmelen, 2004]. We stayed within the description logic portion (OWL-DL) so that we could leverage existing efficient reasoners, such as PELLET[1]. An unexpected outcome of the additional knowledge representation and reasoning was that the same data query workflow is used across the two disciplines. We are finding that it seems 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, Sinha et al. 2007].

Given the value added by this basic knowledge representation and reasoning we extended the method of access to support computer-to-computer interfaces, and in particular via the commonly adopted service oriented architecture implemented as web services.

In Fig.1 at the center toward the top of the figure the primary query entry points are indicated: Instrument, Parameter and Start/Stop Dates. As input to this query workflow are the background ontologies with optional semantic filters: at present to constrain by physical domain (solar physics, middle-upper atmospheric physics, etc.) or an instrument or parameter sub-class (e.g. filter to only query over optical instruments or incoherent scatter radars), the Pellet reasoner and also input from the Metadata Service which, in this case, provides access to the date-time instances due to performance requirements – there are several hundred million instances and instead we 'exit' the ontology and make relational database calls to a mySQL[2] service running on a remote computer. The responses are then re-encoded into OWL for purposes of reasoning.

An example of the web portal interface is shown in Fig. 2. The three query paths, which can be traversed in any order, are to the left and the main response window is immediately to the right of it. The semantic filter options are at the top and the resulting instrument choices are below. Note the partial exposure of the ontology class hierarchy, which allows users to see the lineage of the instrument choices and to date our users like this feature.

1 http://www.mindswap.org/2003/pellet/
2 http://www.mysql.org/

4 Web Services Description and Use

Due to the simplification of the data query workflow as noted in the previous section, the development of web service interfaces naturally followed from the functionality made available in the web portal interface. Thus, in Fig. 1 the three query services are: "by instrument", "by parameter" and "by start/stop date". Also in Fig. 1 to the right, center is the Data Service which provides access to pointers to the data (in our case OPeNDAP[3] URLs). This was the other natural service to expose as a web service and for which we have encoded the inputs, outputs and preconditions in the VSTO ontology[4]. In what follows, we describe each service and give examples of their attributes and use. These examples are driven by use case 2 in section 2.

Fig. 3 shows the example end-point for the Query Instrument semantic web service. In this figure the web service inputs (all optional) and their types are described, and the end-point service address is given along with a link to the Web Services Description Language (WSDL[5]) document content for the service. Two examples are given and below that there is a Query Input form that allows a potential user of the service to scope the types of queries that may wish to make (including the semantic filters discussed earlier).

After the form is submitted a Query Output is returned and displayed on the same page. An excerpt from the corresponding response for the example selection on the form in Fig. 3 is shown in Fig. 4. The returned document is encoded in OWL-DL indicating 13 valid instruments and contains classes and properties and individuals using the VSTO ontology. Such a query results from the same use of the ontology in memory (Jena[6]) and reasoner as would a similar query invoked from the VSTO web portal (e.g. Fig. 2).

A consumer of such a service, either another service, or client application may parse the OWL as XML without semantic meaning (using the background ontology) or directly and use their own reasoning engine (or ours) to further work with the returned information.

3 http://www.opendap.org/
4 http://dataportal.ucar.edu/schemas/vsto_all.owl
5 http://www.w3.org/TR/wsdl
6 http://jena.sourceforge.net

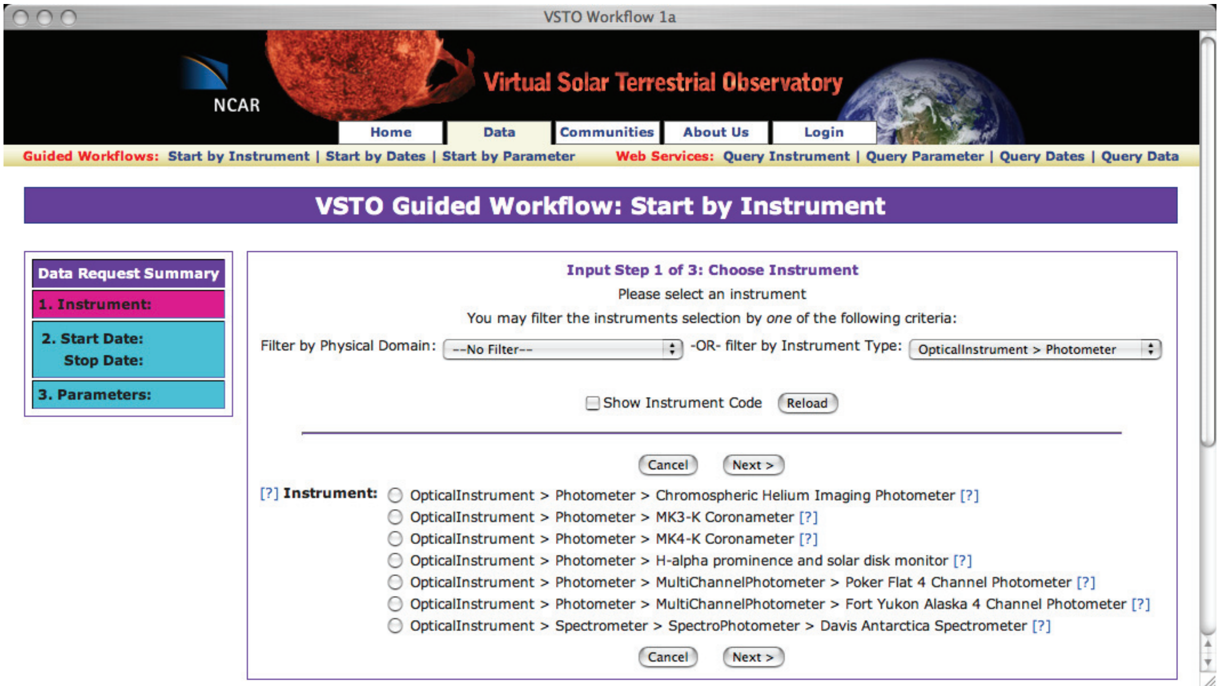

Figure 2: VSTO data search and query interface, exposing taxonomy-based instrument selection.

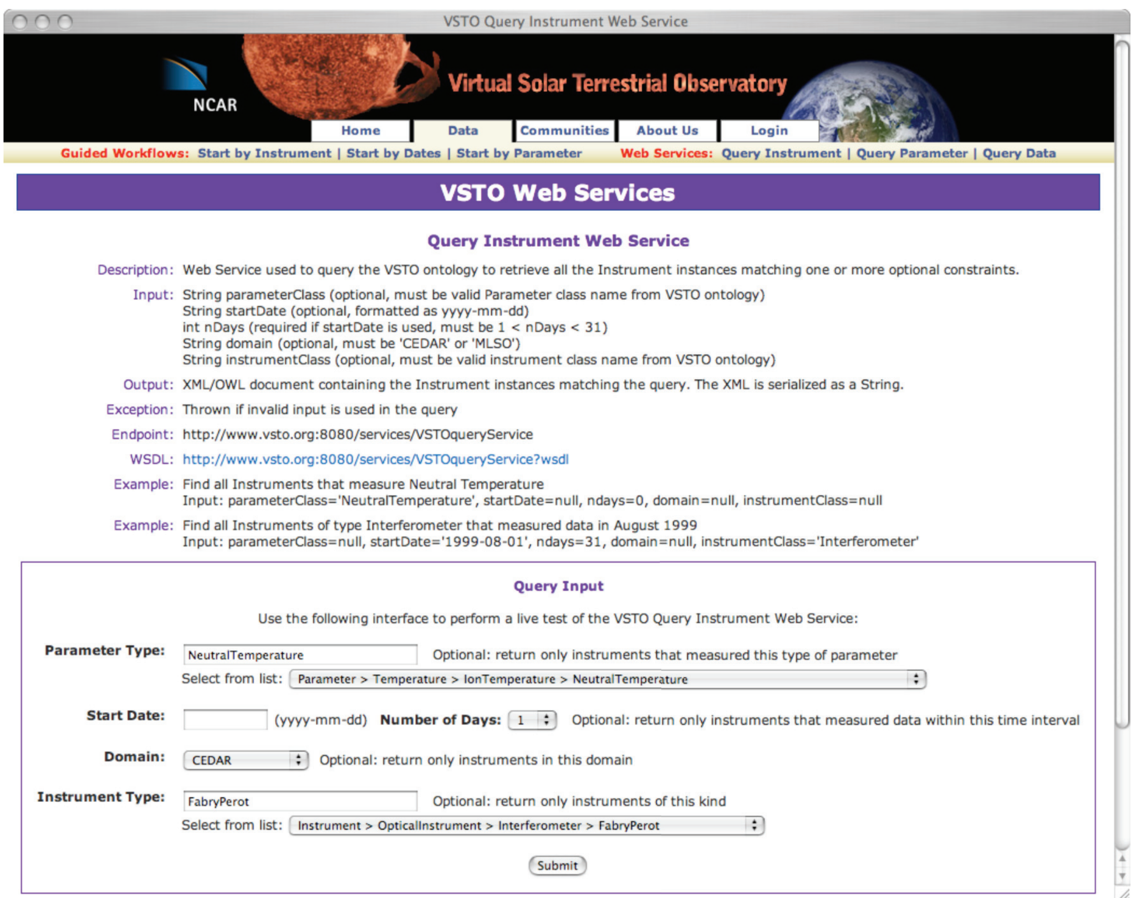

Figure 3: VSTO web services end-point and input example for the query interface initiating an instrument search.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://dataportal.ucar.edu/schemas/vsto_all.owl#"
        xmlns:vsto="http://dataportal.ucar.edu/schemas/vsto.owl"
        xmlns:cedar="http://dataportal.ucar.edu/schemas/cedar.owl#"
        xmlns:mlso="http://dataportal.ucar.edu/schemas/mlso.owl#"
        xmlns:owl="http://www.w3.org/2002/07/owl#" xml:base="http://dataportal.ucar.edu/schemas/vsto_all.owl">
  <vsto:FabryPerot rdf:ID="cedar_instrument_5000">
    <vsto:hasDescription>South Pole Fabry-Perot Interfer Spectr</vsto:hasDescription>
    <vsto:hasName>SPF</vsto:hasName>
    <vsto:hasIdentifier>5000</vsto:hasIdentifier>
  </vsto:FabryPerot>
  <vsto:FabryPerot rdf:ID="cedar_instrument_5015">
    <vsto:hasDescription>Arrival Heights Fabry-Perot Interf Sp</vsto:hasDescription>
    <vsto:hasName>AHF</vsto:hasName>
    <vsto:hasIdentifier>5015</vsto:hasIdentifier>
  </vsto:FabryPerot>
```

Figure 4: VSTO instrument query output excerpt returning OWL documents with semantic information on the available instruments according to the input selections in Fig. 3.

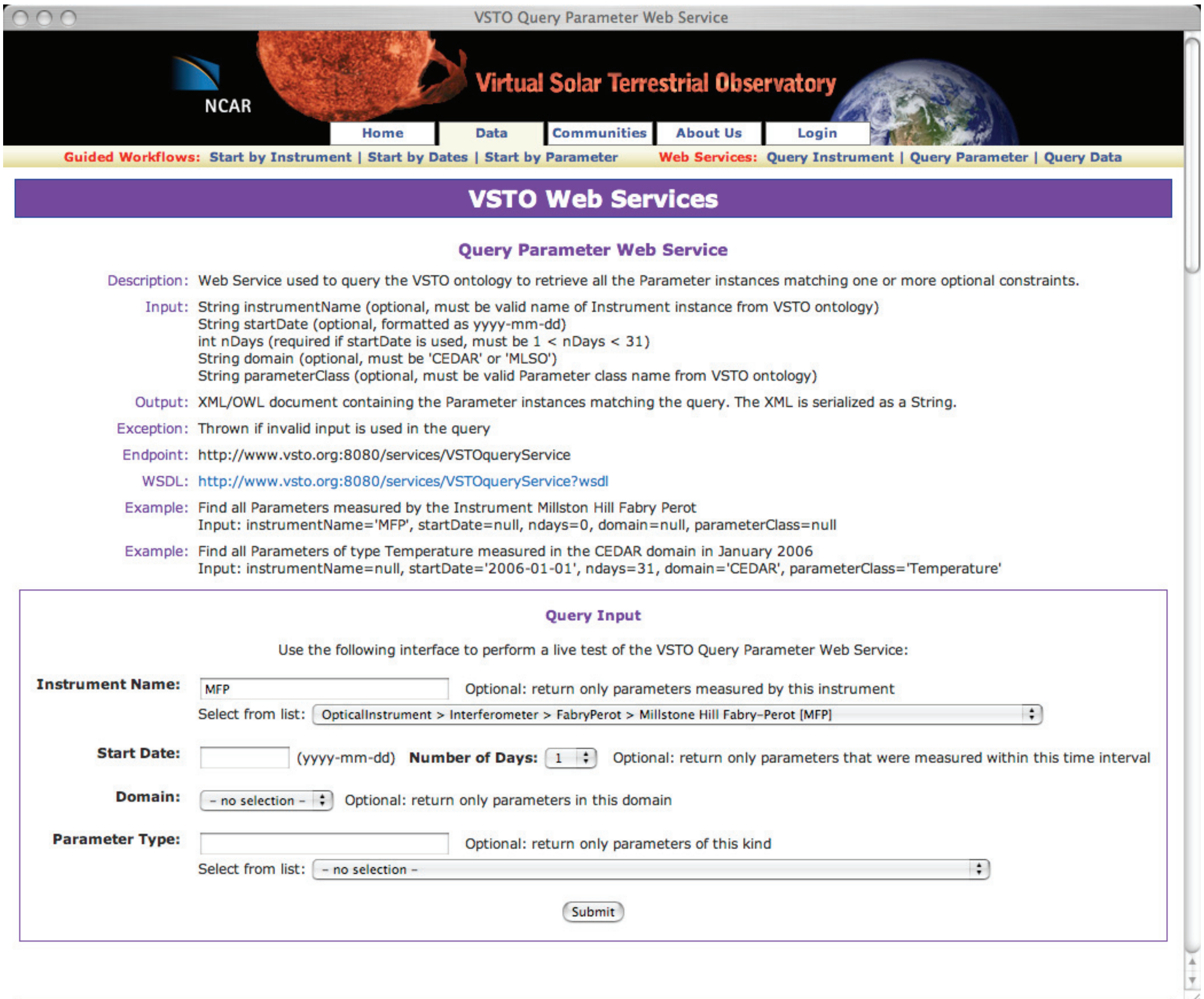

Figure 5: VSTO web services end-point and input example for the query parameter interface initiating a search for parameters matching the query input.

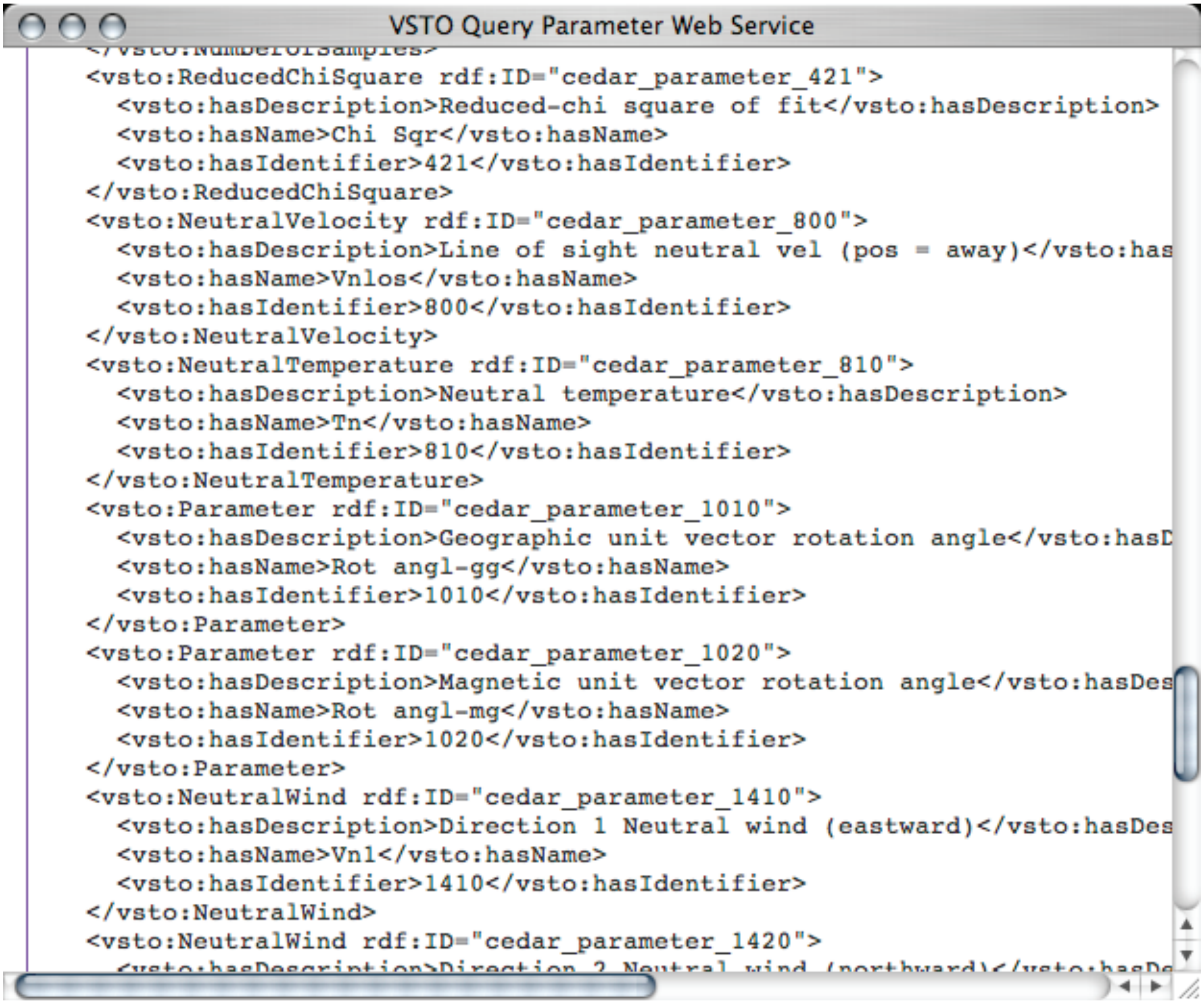

Figure 6: VSTO parameter query search output excerpt returning an OWL document with parameters defined in the VSTO ontology.

Fig. 5 displays a similar example of the end-point for the query parameter service. In this example only one constraint is input to the form selection (the choice of a particular instrument). The Query Output screen, analogous to the query by instruments, is shown in Fig. 6, which indicates 28 returned parameters.

.Figures 5 and 6 together give an example of how services can be chained; for example, a particular (or no) set of constraints can be given to the query instrument service returning a set of instruments and then with a particular instrument choice as input to the query parameter service, a set of valid parameters is returned, and so on. A similar set of chaining is possible using the date-time service as well (not shown here). As a result, clients can combine the three VSTO services in any order and combination to suit their needs.

After suitable queries are made, clients very often want to also return the data associated with particular selections. The VSTO Query Data web service performs this function. Fig. 7 shows the attributes of this service and an example.

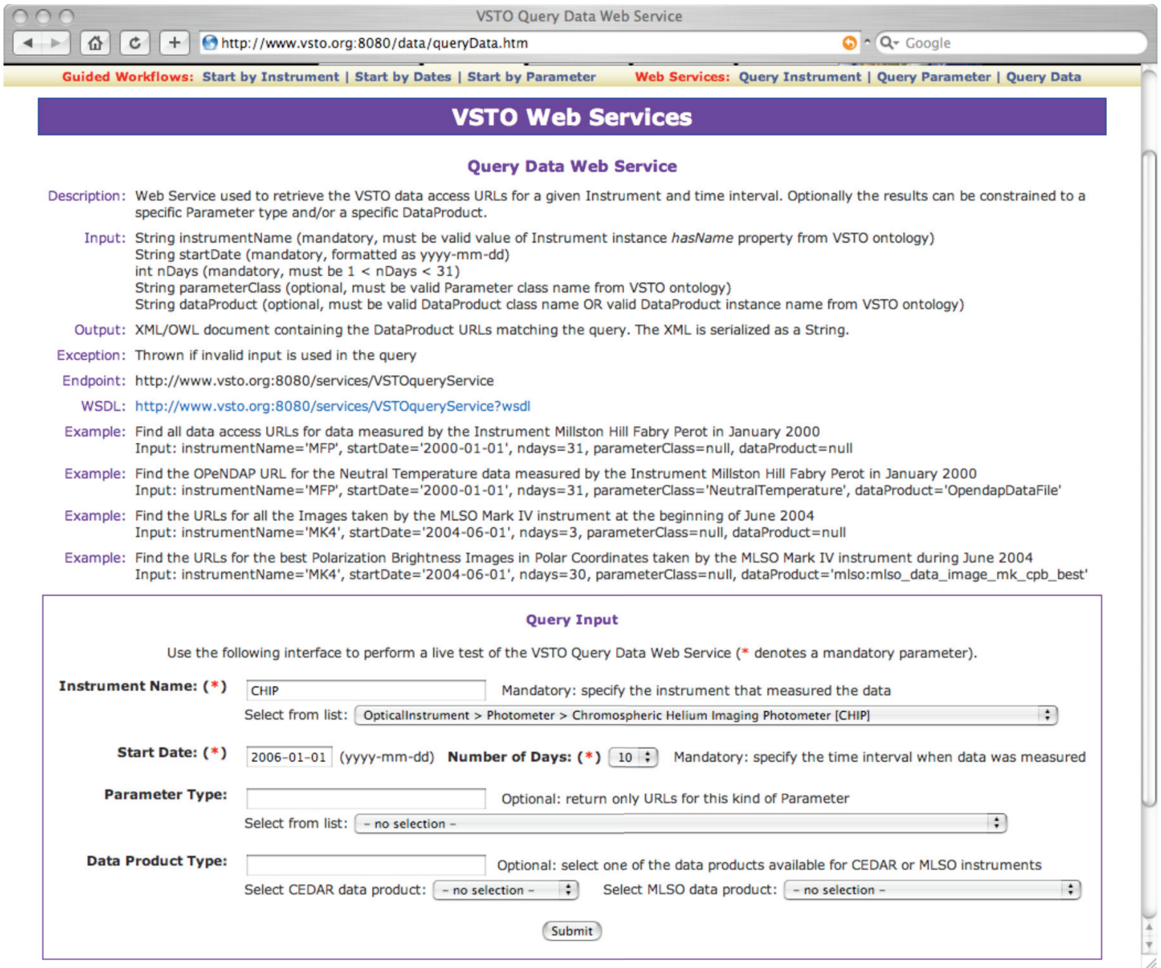

Figure 7: VSTO web services end-point and input example for the data interface initiating a search for links to data matching the query input.

The particular implementation to date has two mandatory inputs: instrument and start/stop dates. The remaining choices of parameter class and data product are optional. While one typical use of the web services interface is to choose ground instances for instruments, a user may choose to use a class description for the instrument, thus allowing more flexibility in retrieval. For example, a user may choose opticalInstrument as a choice in instrument (instead of choosing a particular optical instrument) and then use the web services to discover the optical instruments or to retrieve data from multiple optical instruments (subject to the other constraints in the query). The web services interface thus provides a much greater degree of flexibility for queries. The current portal implementation does not yet include this level of arbitrary use of services

The lower half of Fig. 7 presents an example form to generate OWL output (Fig. 8). In Fig. 8, the first few of the 7000 return results are shown. In this case, the example selects a solar physics instrument that has a variety of data products and data service return types associated with it. For each dataset there are 7 return types (meaning 1000 actual datasets are possible choices) and for each return type the URI/URLs are contained in the OWL document (again, all referencing the VSTO ontology). A client application then either uses these syntactically or semantically leading to one or more de-referencing of the URI/URLs to return data.

The VSTO Data Service web service provides the same type of capability as the SIAP, SSAP, etc. and WxS services discussed earlier except that it offers the semantic encodings for smart use by clients as well as using them in combination, also in a smart way (reasoning, etc.).

Figure 8: VSTO data link query search output excerpt returning OWL documents with links to URLs, which reference the data.

5 Expanding VSTO Portal Functions

At present we have two clients using the VSTO web services; the Virtual Ionosphere-Thermospere-Mesosphere Observatory (VITMO[1]) and the Madrigal Virtual Observatory[2]. Now that these web services are developed and deployed via www.vsto.org, we are now in a position to augment the search and query services we provide in the VSTO portal by installing VSTO web services at remote locations. These services would then be accessed when a user navigates the query workflow (see Fig. 1 and 2) resulting in a distributed set of queries using web services and displayed transparently to the user. Indeed, this is just as a virtual observatory aims to do.

As we integrate with more services, provenance information about such things as data sources used, information recency, authorship (and author credentials) become more critical. We are just beginning our work on representing and providing provenance information. Our initial plans are to encode the provenance information in the provenance fragment of the Proof Markup Language (PML-P) [Pinheiro daSilva et. al., 2006, McGuinness, et.al, 2007]. PML provides an explanation interlingua for use in representing where information came from, how it was manipulated, and any associated trust or reputation information. Once information is encoded in such an interlingua, tools such as those provided by Inference Web [McGuinness, et. al., 2004] may be used to provide users with information about how answers were obtained along with their dependencies and uncertainties. Our initial work focuses on data lineage but future work also includes providing access to information manipulation processes as well.

[1] http://vitmo.jhuapl.edu/
[2] http://madrigal.haystack.mit.edu/madrigal/

6 Summary and Discussion

.

We reviewed our interdisciplinary virtual observatory project – VSTO in its original form as a web portal, which uses semantic technologies to deploy an integrated, virtual repository of scientific data in the fields of solar and solar-terrestrial physics.

We presented the semantic web services we created on the basis of the unified workflow developed as a result of spanning interdisciplinary use cases. As a result we are able to provide simple and meaningful web services to query for data using any chained combination of instruments, parameters, and date-time. Further, we provide a service, which provides access to the URI/URLs directly to the underlying data on interest. All of the web services feature the same knowledge representation and reasoning that is available via the VSTO portal. These services may be used by a client application in both syntactic and semantic form, the latter using the VSTO ontology.

We also discussed how both the provision of web service and the inclusion of remote web service queries can enhance the functionality and coverage of virtual observatories. We have future plans to expose more service functions utilizing the service ontology in addition to or instead of simply returning URI/URLs.

We also have plans to evolve our ontologies and infrastructure as required to cover broader science areas. We have reviewed the basic structure of the VSTO ontologies with respect to the needs of related science projects in related domains including the NSF-funded GEON project, the NASA-funded SESDI project, and the NASA-funded SKIF project. We have found that the ontology structure, the method of query workflow selection and our services carry over to these rather distinct discipline areas (with additions to the ontology to reflect both domain content and instrument./parameters of interest for those fields). .

To provide even greater semantic value, 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, as we do these updates, we will be adding provenance information to the existing and expanded content so we can provide access on demand to where the information came from.

Acknowledgements

The VSTO project is funded by the National Science Foundation, Office of Cyber Infrastructure under the SEI+II program, grant number 0431153. The National Center for Atmospheric Research is operated by the University Corporation for Atmospheric Research with substantial sponsorship from the National Science Foundation. The authors which to acknowledge useful comments from the reviewers.

References

Bentley, R., Bogart, R., Davis, A., Hurlburt, N., Mukherjee, J., Rezapkin, V., Roberts, D.A., Szabo, A., Weiss, M. 2005, A Framework for Space and Solar Physics Virtual Observatories, hpde.gsfc.nasa.gov/VO_Framework_7_Jan_05.pdf

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.

Deborah McGuinness, Peter Fox, Luca Cinquini, Patrick West, Jose Garcia, James L. Benedict, and Don Middleton. The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research. In the proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07). Vancouver, British Columbia, Canada, July 22-26, 2007.

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.

Sinha, K, McGuinness, D.L., Fox, P.A., Raskin, R., Condie, K., Stern, R., Hanan, B and Seber, D. 2007. Towards a Reference Plate Tectonics and Volcano Ontology for Semantic Scientific Data Integration, U.S.Geological Survey Scientific Investigations Report 2007, in press.