SPASE AND THE HELIOPHYSICS VIRTUAL OBSERVATORIES

J. R. Thieman[1]*, D. A. Roberts[2], T. A. King[3], C. C. Harvey[4], C. H. Perry[5], and P. J. Richards[5]

[1]**Code 690.1, NASA/GSFC, Greenbelt, MD 20771, USA*
Email: james.r.thieman@nasa.gov
[2]*Code 672, NASA/GSFC, Greenbelt, MD 20771, USA*
Email: aaron.roberts@nasa.gov
[3]*UCLA, Institute of Geophys. and Planetary Phys., University of California, 5881 Slichter Hall, Los Angeles, CA 90095-1567, USA*
Email: tking@igpp.ucla.edu
[4]*CESR -CNRS, B.P. 4346, 9, avenue du Colonel Roche, 31028 Toulouse CEDEX 4, France*
Email: harvey@cesr.fr
[5]*Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, United Kingdom*
Email: c.h.perry@rl.ac.uk
Email: p.j.richards@rl.ac.uk

ABSTRACT

The Space Physics Archive Search and Extract (SPASE) project has developed an information model for interoperable access and retrieval of data within the Heliophysics (also known as space and solar physics) science community. The diversity of science data archives within this community has led to the establishment of many virtual observatories to coordinate the data pathways within Heliophysics subdisciplines, such as magnetospheres, waves, radiation belts, etc. The SPASE information model provides a semantic layer and common language for data descriptions so that searches might be made across the whole of the heliophysics data environment, especially through the virtual observatories.

Keywords: Data, Metadata, Model, SPASE, Observatory, Information, Heliophysics, Interoperability

1 INTRODUCTION

The Heliophysics (space and solar physics) data environment consists of many data sets from both satellite and ground-based instruments measuring our dynamic Sun and the effects that it has on the solar system, especially the Earth. This involves not only the light and heat radiation given off by the Sun but electromagnetic radiation of all wavelengths, as well as the charged and neutral particles that make up the solar wind. The effects can be observed through electric and magnetic field measurements, particle and wave spectroscopy, energy determinations, aurora monitoring, etc. Series of ground-based measurements have been ongoing for over a hundred years, and Heliophysics experiments have been carried on the majority of satellites since the dawn of the space age. Consequently, the Heliophysics data environment is complicated by many data sets archived in many places around the world. Older data sets tend to be relatively small, but the latest satellite and ground-based instruments can have associated data sets that are quite voluminous even by today's standards.

2 INTEROPERABILITY PROBLEMS

Heliophysics research often requires input and intercomparison of many data sets. The scientist needs to know where the needed data reside and how to access and use them. Many years ago the Heliophysics community was smaller, and it was not hard to know where most of the data sets could be found or who had the knowledge of what was available. As the number of spacefaring nations increases, it becomes much more difficult to keep track of all potentially useful data. SPASE and the Virtual Observatories were developed to help solve this problem.

2.1 Heliophysics vs. Astrophysics

In the Astrophysics community the proliferation of data has been a problem far longer than in Heliophysics. The number of ground-based and space-based telescopes contributing research quality data is large and growing. Yet the

Astrophysics data environment is simpler in many ways than Heliophysics.

• Astrophysics researchers usually search for an object, a particular sky location, and/or a particular wavelength for the data they need, whereas Heliophysicists need to search by time, location, object, measurement type, energy, and many other factors.
• Astrophysics space-based data sets and many of the ground-based data sets as well tend to be large in volume and made up mainly of imagery, but Heliophysics has a multiplicity of small data sets that measure fields and particles in a large variety of ways including imagery.
• Because Astrophysics data is dominated by imagery, their research community has agreed to relatively few standard formats for data with FITS and VOTable formats being dominant; the Heliophysics community does not have a dominant format and the many types of data result in a large variety of formats.
• The limited number of standard formats in Astrophysics makes it much easier to develop standard metadata terminology in comparison to Heliophysics where the terminology has long been quite diverse.
• Astrophysics data archives can often be classified according to the wavelength of the data, but Heliophysics data must be grouped by a variety of data types.

Astrophysics data can be quite complex with spectral data, time series data, catalogs, etc. but these have been formatted to use the FITS standard. It is not a simple matter to make these data interoperable, but the problem has been addressed for many years, and agreement on standards is more advanced than in Heliophysics.

2.2 Heliophysics Interoperability

Heliophysics is only now establishing a common approach to making the data readily retrievable and interoperable and still has great difficulties getting agreement for common ways to handle data and metadata. It is necessary for representatives from a wide variety of Heliophysics subdisciplines to come together and undergo prolonged discussions to agree on a common approach that will work for that community. When a common approach can be found to describe the variety of data in the discipline archives and allows them to be easily intercompared, then "interoperability" within that data environment is enhanced. Thus, Astrophysics has a well-established Virtual Observatory approach to finding, retrieving, and intercomparing data as the groundwork has been done to make the data more readily interoperable.

3   HELIOPHYSICS VIRTUAL OBSERVATORIES

A virtual observatory can be defined as one or more software systems and tools connected to data centers and archives through the internet in order to enable the rapid location and retrieval of useful data for scientific research. "The idea of the Virtual Observatory (VO) is to achieve the same transparency for astronomical data. In the VO all the world's data is available from your desktop. All archives understand the same query language, can be accessed through a uniform interface, and diverse data can be analyzed by the same tools (Quinn et al., 2004)." Note the reference to "astronomical" data. The virtual observatory concept has been synonymous with the astronomical community, but the approach can be used in any reasonably cohesive data environment.

Within Heliophysics, virtual observatories have been established with the purpose of providing access to various subdisciplines within the community. Many of these have been formed with funding from the National Aeronautics and Space Administration (NASA). The present NASA-funded VOs are:

• VSO - Virtual Solar Observatory
• VSPO - Virtual Space Physics Observatory
• VMO - Virtual Magnetospheric Observatory
• VITMO - Virtual Ionosphere, Thermosphere, Mesosphere Observatory
• VHO - Virtual Heliophysics Observatory
• ViRBO - Virtual Radiation Belt Observatory
• VEPO - Virtual Energetic Particle Observatory
• VWO - Virtual Wave Observatory
• VMR - Virtual Model Repository

The VSO was the first of the VOs to be formed. The others have been created more recently as a result of competitive calls for proposals for these observatories. They were chosen according to the needs within the subdisciplines and what would best fulfill those needs. They are governed and coordinated through the Heliophysics Data and Model

Consortium (HDMC), a NASA project that is part of their Heliophysics data system infrastructure, formed, in part, to deal with the problem of integrating the VxOs. Progress has been steady through this approach, but it is still a challenge to keep the number of diverse systems working together to a common goal. This will be a focus of the next two years of effort. More and more of the available resources are being described using SPASE and integrated into the inventories of the Virtual Observatories. Work is underway to define services and protocols to achieve full interoperability among NASA's Virtual Observatories and with those of international partners

Although these data systems were founded specifically as virtual observatories, there are other facilities within the community that are not NASA-funded and can be considered to fulfill the role of a virtual observatory. These include the following:

- CAA - Cluster Active Archive
- CDPP - Centre de Données de la Physique des Plasmas
- CSSDP - Canadian Space Science Data Portal
- EGSO - European Grid of Solar Observations
- GAIA - Global Auroral Imaging Access
- VSTO - Virtual Solar Terrestrial Observatory

Work is underway to define services and protocols to achieve full interoperability among NASA's Virtual Observatories and those of international partners.

4   THE SPASE DATA MODEL

As might be expected, the many virtual observatories in Heliophysics are often quite different from each other, and there are still important data centers that are not directly associated with a virtual observatory. In order to make data "findable" in this environment, it is necessary to have a common metadata language which is used to describe the data available through all of the virtual observatories, data centers, and archives in a common way. The Space Physics Search and Extract (SPASE) project is an international collaboration of representatives of the Heliophysics data community working together to provide common search and data retrieval capabilities based especially on a SPASE data model.

The SPASE Data Model has been developed through several years of interaction of the SPASE Working Group mainly through biweekly teleconferences and email exchanges.

The working group has several goals:
- Facilitating data search and retrieval across the Space and Solar Physics data environment;
- Defining and maintaining a standard data model for Space and Solar Physics interoperability;
- Demonstrating the Model's viability;
- Providing tools and services to assist SPASE users; and
- Working with other groups for other Heliophysics data management and services coordination as needed.

As far as is known, the SPASE Working Group is the only international group attempting to achieve interoperable global data management for Solar and Space Physics. The group is self-formed but had received funding for some of the Virtual Observatory personnel to participate. This funding has ended, but it is understood that the NASA-supported Virtual Observatories will adopt SPASE as their *interlingua* for communication between the observatories. The number of other U.S. and international participants in the SPASE effort is growing even without the impetus of funding. A list of the those who have participated in the SPASE working group is far too large to include here, but a list of the participants may be found at the beginning of the SPASE Data Model document (pages I and II) downloadable from the SPASE website (http://spase.gsfc.nasa.gov).

The SPASE Data Model will only be described briefly here.   The document defining the SPASE Data Model as well as supporting information about it may be found at the SPASE web site. The model is broken down into a number of main elements known as Resource Types, each defining a main topic of data description. The Resource Types are presently:

Annotation   Explanatory or descriptive information that is associated with another resource.

Catalog	Listing of events or observational notes.
Document	Plain or formatted text that may include graphics, sound, other multimedia data, or hypermedia references.
Display Data	A graphical representation of data.
Numerical Data	Data product stored as numerical values in a specified format.
Granule	An accessible portion of another resource.
Instrument	A device that is used to sense and parameterize a physical phenomenon.
Observatory	The host (spacecraft, network, facility) for instruments making observations.
Person	An individual human being.
Registry	A location or facility where resources are catalogued.
Repository	A location or facility where resources are stored.
Service	A location or facility that can perform a well-defined task.

Each of these elements is broken down into sub-elements that provide sufficient detail for adequate description of data sets. The Resource Types Numerical Data, Display Data and Catalogue are the resources used to describe Heliophysics data products. These data product descriptions reference the other resources which contain descriptions of the observatories, instruments, people, etc. that created the data products. The resources are described using the terms from the SPASE Data Dictionary presented in the Data Model document.

Presently there are several versions of the data model that appear on the web site, and these reflect the progression of honing the model based upon feedback from those using it to describe their data. The official version at the time of this writing is version 2.0. To get a feeling of the hierarchy of the model, one good approach is to click on "Tree" under the Data Dictionary on the SPASE web site home page. This displays the model terminology in a hierarchical format indicated by indentation, and the terms themselves are linked to their definitions through a simple click. Figure 1 illustrates one part of the tree as displayable on the web site.
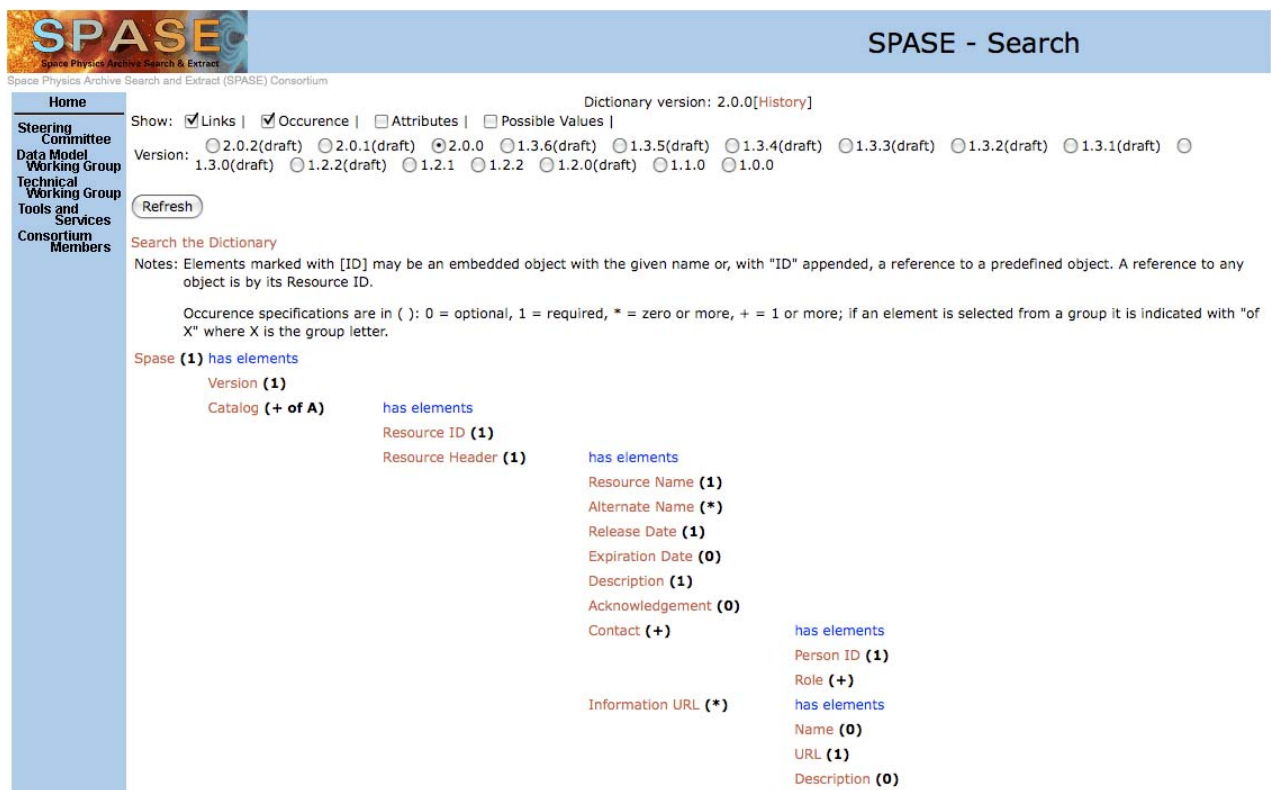

Figure 1. A portion of the SPASE Data Model hierarchy displayed in "tree" format on the SPASE web site

A number of application tools have been developed to facilitate the use of the SPASE Data Model. These are listed on the SPASE web site and accessible through that site. The types of tools listed there include: validator, parser, editor, generator, harvester, wrapper, etc. There is not enough space to describe all of these, so it is left to the interested reader to access the information on the web site.

5   SEARCH SCENARIO

The proof of the SPASE Data Model will be in its usage for searching and retrieving data in the Heliophysics data environment. Let us look at a research scenario and determine how the virtual observatory approach works now and how it might work in the future.

Assume a researcher is interested in finding a wide variety of data relevant to a study of the major solar storms from Oct. 31, 2003, commonly known in the U.S. as the "Halloween Storm" due to the celebration that occurs on that date. The time span of the search is from 2003-10-30 00:00:00 to 2003-11-01 23:59:59. To search across all archives one-by-one is impractical given the number of archives of interest, probably as many as 100. The virtual observatories span a number of these archives so that is one approach that is more feasible. We will try this approach with a few sample virtual observatories and then consider a SPASE–related search across multiple VOs.

The first example is the Virtual Solar Observatory (VSO - http://sdac.virtualsolar.org/cgi-bin/search). This was the first Heliophysics virtual observatory developed, and it predates the SPASE effort. The interface it uses for searching assumes a reasonable knowledge of solar physics observables. An image of the VSO search interface is shown in Figure 2. If the search in VSO is done with the indicated time frame and all available observables chosen, there are at the time of this writing 6695 data entries returned having both narrow and wide time ranges. Obviously this search would need to be narrowed in some way, but it is indicative of the amount of data that might be accessed through a VO search. A method of searching VSO via SPASE terminology will be provided in the future.

The Virtual Ionosphere, Thermosphere, Mesosphere Observatory (VITMO - http://mizar.jhuapl.edu/sras/frameset.jsp) has a very different approach to data searching. As shown in Figure 3, the search interface has a relatively simple appearance, which hides a sophisticated database that underlies the search mechanism. VITMO incorporates a "discovery" type of approach to data searching. When building a query through this interface, a temporal range can be specified, but there is also a category of user-defined events, and one of the events is defined as "Halloween Storm October 2003." In selecting this event, the temporal range is automatically set to 2003-10-24 00:00:00 to 2003-11-01 23:59:59. This can be modified, however, to our period of interest starting on Oct. 30, 2003. Some parameters must be chosen to get a valid search. In this case all of the parameters associated with the ionosphere region were chosen, and the result was 21 products from ROCSAT (Republic of China Satellite), SuperDARN (Super Dual Auroral Radar Network), and the TIMED (Thermosphere Ionosphere Mesosphere Energetics and Dynamics) mission. The number of individual data results is not initially indicated. Again, the terminology being used within the interface is being mapped to the SPASE terminology for use in the future.
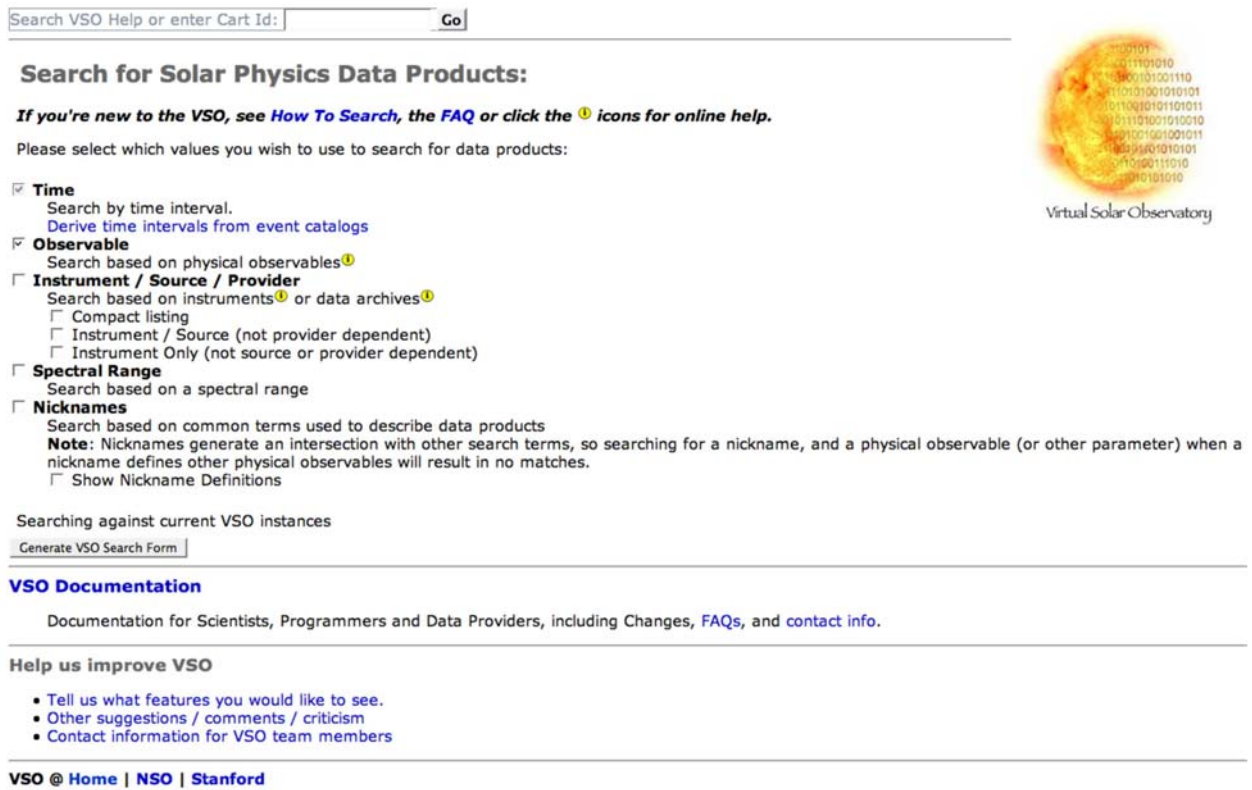
Figure 2. The Virtual Solar Observatory search interface starting page.
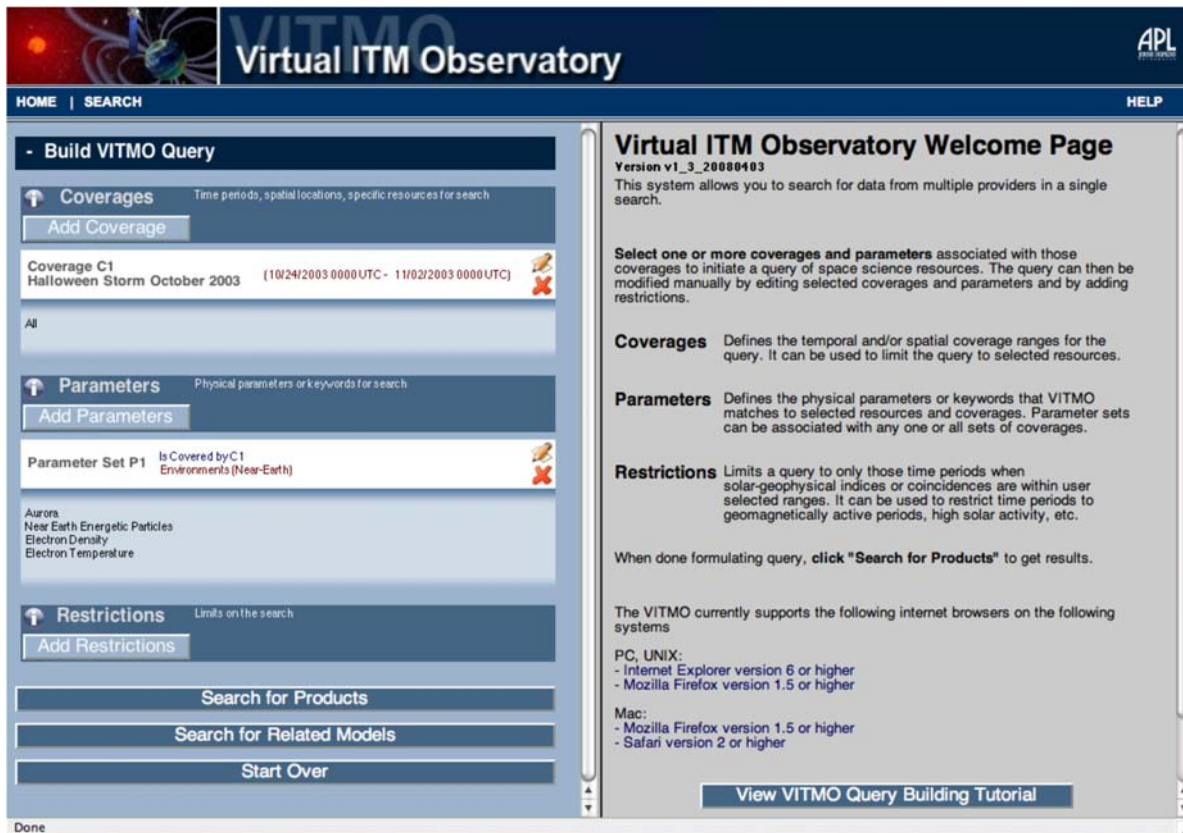

Figure 3. The Virtual Ionosphere Thermosphere Mesosphere Observatory search interface starting page.

The Virtual Magnetospheric Observatory offers several search approaches, but one that uses a hierarchical query builder type of search based on SPASE terminology as shown in Figure 4 (VMO - http://vmo.nasa.gov/index.php/data-query-mainmenu-93/gsfc-interface-mainmenu-70). In this case, using a time search yields 155 data files at the time of this writing. Many of these are ground-based magnetometer files.

Finally, the Virtual Space Physics Observatory (VSPO - http://vspo.gsfc.nasa.gov/websearch/dispatcher) is built to use the SPASE data descriptions directly as they are created in XML. VSPO provides access not only to specific space physics data sets but also can be used to access all discipline areas by importing all of the SPASE data descriptions that are created. That is not completely done at the time of this writing but can be done in the future and will be used as one method to search across all of the data described through the SPASE Data Model. With the present data holdings in VSPO, a simple time search for the time period covering the Halloween Storms yields 244 data products. A "get data" button is offered for many of these products fulfilling the need not only to find useful data but extract the data as well. Figure 5 shows the VSPO search interface as with just the time search invoked although the other SPASE resources are listed on the left hand side for further narrowing of the query if needed. The beginning of the list of results is shown in the columns on the right hand side.

Figure 4. The Virtual Magnetospheric Observatory search interface starting page.

Figure 5. The Virtual Space Physics Observatory search interface starting page.

6   STATUS

Although there is no cross-VO SPASE search approach yet implemented, any VO can gather all of the SPASE data descriptions and use them for searches according to what is of interest to the users. VSPO is intended to do this. The question is whether there will ultimately be an interface that searches across all of the VOs simultaneously and translates results into a common form. This could also include searches of data archives that may not be associated with any particular VO.

Development of the SPASE Data Model continues in biweekly teleconferences and occasional face-to-face meetings. Data descriptions continue to be generated according to the Data Model, and feedback from problems associated with creating these data descriptions is factored into the further development of the model. The teleconferences are used to resolve the problems in a reasonably timely way. Presently, version 2.0 is the official version of the model, but new developments are being incorporated, and it is anticipated that soon the changes will be locked in with the release of 2.1. Support for the continued development of SPASE comes from NASA Headquarters, and the central coordination point of the SPASE effort is within the National Space Science Data Center (NSSDC) at Goddard Space Flight Center. Funding for SPASE is now a part of the NSSDC budget and will continue on into the foreseeable future.

7   EFFECTIVENESS

As SPASE continues to be developed the effectiveness of the SPASE data model is continually evaluated. The assessment asks the following questions:

• How well does SPASE function as an *interlingua* among the Virtual Observatories and data archives?

• How effective is SPASE in describing data sets for data finding and usage?

• How much should SPASE be "inside" vs. "outside" the observatories, etc. to be effective?

- To what level of detail should data descriptions be created in order to fulfill the objectives of SPASE and the Space Physics data environment?

- What else is needed in SPASE for the non-NASA virtual observatories environment?

- What interfaces are most effective for SPASE searches?

- Should SPASE be expanded for usage in closely-related disciplines such as Planetary Science?

Answers to these issues will not be immediate. These will be determined through trial and error and user feedback as the project continues

8   CONCLUSION

The utility of the Virtual Observatories and the SPASE Data Model is currently being demonstrated in NASA's Heliophysics Data Environment. Participation of the Heliophysics science community is important to assuring the approach is both effective and matures to meet the needs of researchers. SPASE data descriptions are being created by several groups, but it would help to have others involved in this process. This tests not only the utility of the SPASE Data Model but also the tools that have been developed to facilitate the description of data and the usage of the model. Feedback from all who are involved is critical to future success!

9   ACKNOWLEDGEMENTS

10   REFERENCE

Quinn, P., Lawrence, A., & Hanisch, B. (2004) A discussion paper for the OECD global science forum, *International Virtual Observatory Alliance Website.* Retrieved from the World Wide Web, February 16, 2010: http://ivoa.net/pub/info/.