The SPASE Data Model for Heliophysics Data: Is It Working?

James Thieman [1], Todd King [2], and Aaron Roberts [3]

[1] ***NASA/GSFC***
Code 690.1, NASA/GSFC, Greenbelt, MD 20771, USA
EMail: james.r.thieman@nasa.gov

[2] ***UCLA/IGPP***
Institute of Geophysics and Planetary Physics, University of California Los Angeles, CA 90095, USA
EMail: tking@igpp.ucla.edu

[3] NASA/GSFC
Code 672, NASA/GSFC, Greenbelt, MD 20771, USA
EMail: aaron.roberts@nasa.gov

ABSTRACT

The SPASE (Space Physics Archive Search and Extract) Data Model was developed to provide a metadata standard for describing Heliophysics (Space and Solar Physics) data within that science discipline. The SPASE Data Model has matured over the many years of its creation and is presently represented by Version 2.2.1. Information about SPASE can be obtained from the website http://spase-group.org. The Data Model defines terms and values as well as the relationships between them in order to describe the data resources in the Heliophysics data environment. This data environment is quite complex, consisting of Virtual Observatories, Resident Archives, Data Providers, Partnering Data Centers, Services, Final Archives, and a Deep Archive. SPASE is the metadata language standard intended to permeate the complexity and provide a common method of obtaining and understanding data. Is it working in this capacity?

SPASE has been used to describe a wide range of data. Examples range from ground-based magnetometer data to interplanetary satellite measurements to space weather model results. Has it achieved the goal of making the data easier to find and use? To find data of interest it is necessary that all the data of importance be described using the SPASE Data Model. Within the part of the data community associated with NASA (supported through NASA funding) there are obligations to use SPASE and to describe the old and new data using the SPASE XML schema. Although this part of the community is not near 100% compliance with the mandate, there is good progress being made and the goal should be reachable in the future. Outside of the NASA data community there is still work to be done to convince the international community that SPASE descriptions are worth the cost of their generation. Some of these groups such as Cluster, HELIO, GAIA, NOAA/NGDC, CSSDP, VSTO, SuperMAG, and IUGONET have agreed to use SPASE, but there are still other groups of importance that need to be reached. It is also assumed that the terminology is sufficiently broad and the descriptions are sufficiently complete that researchers needing data of a specific type or from a specific period can find and acquire what they need. A valid SPASE description can be very brief or very thorough depending on the willingness of the author to spend the time necessary to make the description useful. There is evidence that users are finding what they need through the SPASE descriptions, and this standard is a big step forward in Heliophysics data location.

Does SPASE make it easier to use the data once they are found? Thorough descriptions of data using SPASE can describe the data down to the level of individual parameters and exactly how the data are organized and stored. Should the SPASE data descriptions be written in such a way that they can be automatically ingested and understood by software tools? Heliophysics instruments are becoming more versatile all the time and the complexity of the data makes it tedious and time consuming to write SPASE descriptions with this level of sophistication even with the improvement of the tools used to generate the descriptions. Is it better to just write human-readable descriptions of the data at the parameter level or to refer to references that provide this information? This is a debate that is presently taking place and software is being developed to test what is possible.

Keywords: up to 8 keywords, comma separated, may be added here

INTRODUCTION

The Space Physics Archive Search and Extract (SPASE) project has been discussed in several of the past PV meetings. The next section will again give a brief overview of the project and the purpose behind it and then in the following sections the latest progress of the effort will be reviewed. Finally, there will be a section describing the plans for the future and the steps that will lead to the anticipated outcome.

The solar and space physics community (usually called the Heliophysics discipline within NASA) is quite diverse and does not have any overall governing body. There are loose ties among the organizations that contribute data to the general study of the Sun's variability, space weather, and the effects on the Earth and the other parts of the solar system. There are a large variety of sets of data that contribute to this overall study, both from satellites and ground-based instruments. Figure 1 indicates the complexity of just the current satellite-based aspect of the Heliophysics data environment.

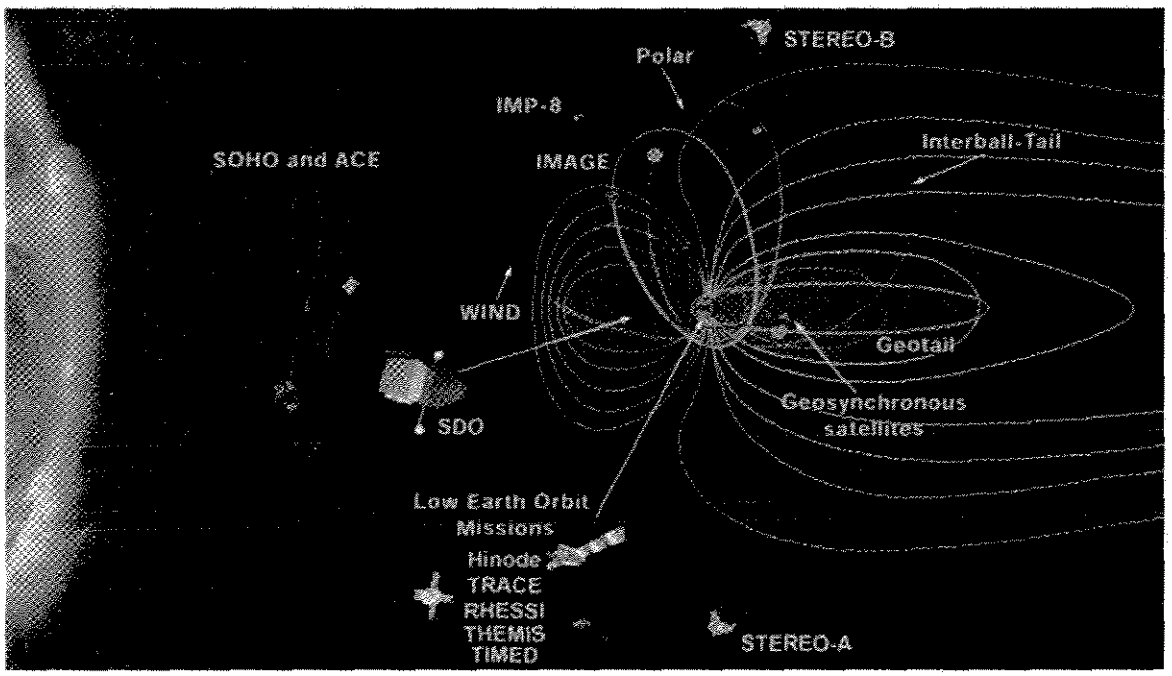

Figure 1. The Heliophysics "Great Observatory" of satellites supporting the study of solar and space physics. Some other satellites and all ground-based instruments are not included in the figure.

Although Figure 1 is relatively complete in terms of the number of satellites contributing data to the Heliophysics data archives today, there are others that could also be added as well as the many ground-based observatories and instruments that add to the knowledge base of the factors that affect solar and space physics understanding.

SPASE AND THE HELIOPHYSICS DATA ENVIRONMENT

The Heliophysics Data Environment is rather complex as well. It has evolved as a loose conglomeration of many different data archives storing the data from the satellites and ground-based instruments, as well as model-generated data and services groups. Figure 2 gives an idea of the diversity of this data environment mainly from the viewpoint of the NASA-supported aspects of the data network. There are

quite a number of acronyms in this Figure, but the list below the figure provides the interpretation for them.

A Heliophysics researcher looking for useful data within the NASA-related data archives in the environment should have some idea of which subdisciplines of space and solar physics might have data of use for the intended research. If the research is concentrated within a particular subdiscipline then the researcher could go directly to the virtual observatory associated with that subdiscipline. If, however, the research cuts across subdisciplines then it becomes more difficult to contact all of the systems that may have data of interest. This task is further complicated by the many other data providers, international and NSF partners, final archives, and the deep archive, that might also have needed data. One might also want to know about the many services and tools that are available on the network to help analyze the data. Practically every system uses a different approach to describe the data that they hold or the services they provide.

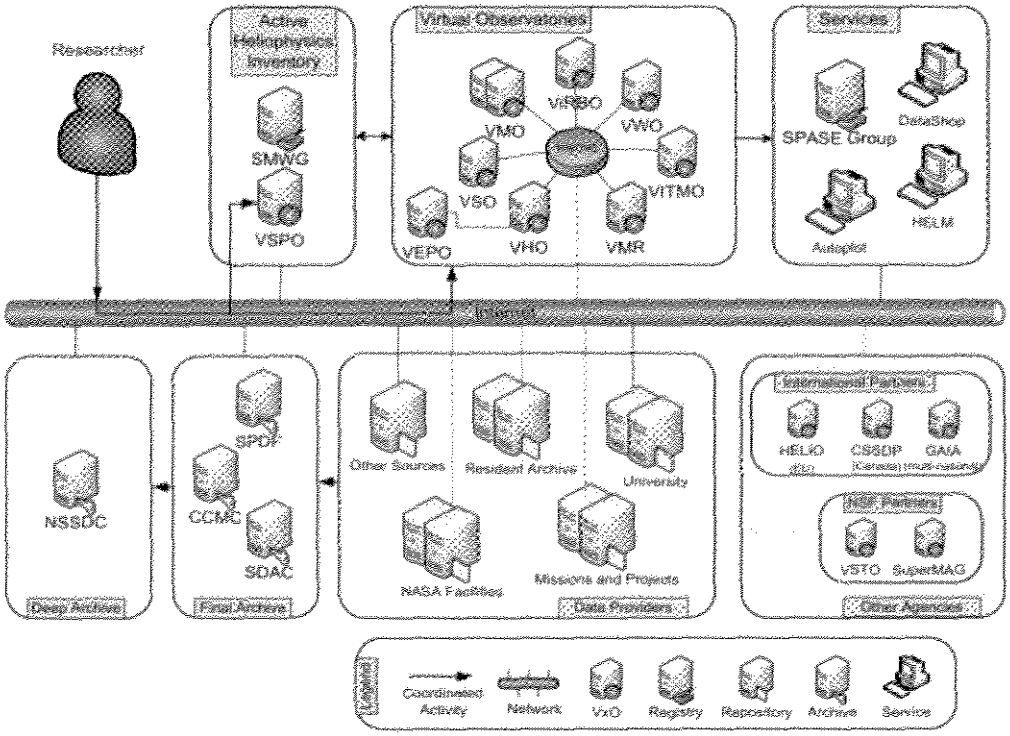

Figure 2. The Heliophysics Data Environment is a rather complicated interconnection of NASA-sponsored and general community systems and services. The acronyms are listed in Table 1.

It would be nice to have some tool that would allow searches of the entire data environment for topics of interest. Then it is usually better to use the active inventories based upon this tool to determine where the data might reside. This is the purpose behind the development of the SPASE Data Model. A consortium of the organizations involved in the data environment have representatives participating in the SPASE project. This group has developed a SPASE Data Model as a common format to describe the data available through the data environment. Assuming that all the data sets of interest are described using the SPASE Data Model then common terminology can be used to search all of the data descriptions and point the way to the facilities that contain the data. The common language can also be used to request data from the archives and to download the data together with sufficient description for efficient use in data analysis. Thus, the SPASE data model provides a common translational layer on top of the existing data

descriptions to make all the systems interoperable. The extent of interoperability depends on the level of support built into the SPASE Data Model and the care put into the SPASE data description. To learn more about the SPASE Data Model it is best to download the document describing the latest version of this Model. Currently the latest version is Version 2.2.1 and it may be downloaded by going to http://spase-group.org and clicking on the "Current Version" link on the upper right hand side. The document describes the XML-based schema used to describe the data sets.

CCMC	Community Coordinated Modeling Center
CSSDP	Canadian Space Science Data Portal
GAIA	Global Auroral Imaging Access
HELIO	Heliophysics Integrated Observatory
HELM	Heliophysics Event List Manager
NSSDC	National Space Science Data Center
SDAC	Solar Data Analysis Center
SMWG	Science Metadata Working Group
SPASE	Space Physics Archive Search and Extract
SPDF	Space Physics Data Facility
SuperMAG	The Global Ground-Based Magnetometer Initiative
VEPO	Virtual Energetic Particle Observatory
VHO	Virtual Heliophysics Observatory
VIRBO	Virtual Radiation Belt Observatory
VITMO	Virtual Ionosphere, Thermosphere, Mesosphere Observatory
VMO	Virtual Magnetospheric Observatory
VMR	Virtual Model Repository
VSO	Virtual Solar Observatory
VSPO	Virtual Space Physics Observatory
VSTO	Virtual Solar Terrestrial Observatory
VWO	Virtual Wave Observatory

Table 1. Acronym list for Figure 1.

Figure 3 gives a high level overview of the present state of information supported in the SPASE Data Model. The highest categories of information are called "resources" and there are four basic types of resources: Data, Entity, Granule, and Person. Both Data and Entity include subcategories of information that contain provenance information about them. The "Data" Resource is divided into a number of types such as numerical data, display data, catalog data, etc. The general category of "Entity" provides a place for the main metadata associated with the data such as the related observatory, instrument name, associated repository, etc. Another resource is any person to be associated with the data. This is usually the cognizant contact person, but it could also be the principal investigator, a technical expert, etc. Finally, there is the resource called "Granule" which represents a subset of the overall data set. The description of the overall data set also applies to the granule, but then it has additional information indicating the factors that determine the uniqueness of the subset. The SPASE project is also considering whether "Granule" should have metadata attributes sufficient to allow data analysis such as plotting of the

data to be performed automatically. This has not been done yet and some argue that this should be information that is packaged with the data or obtained via a separate route. At the moment SPASE contains only the minimal amount of support for description of data granules.

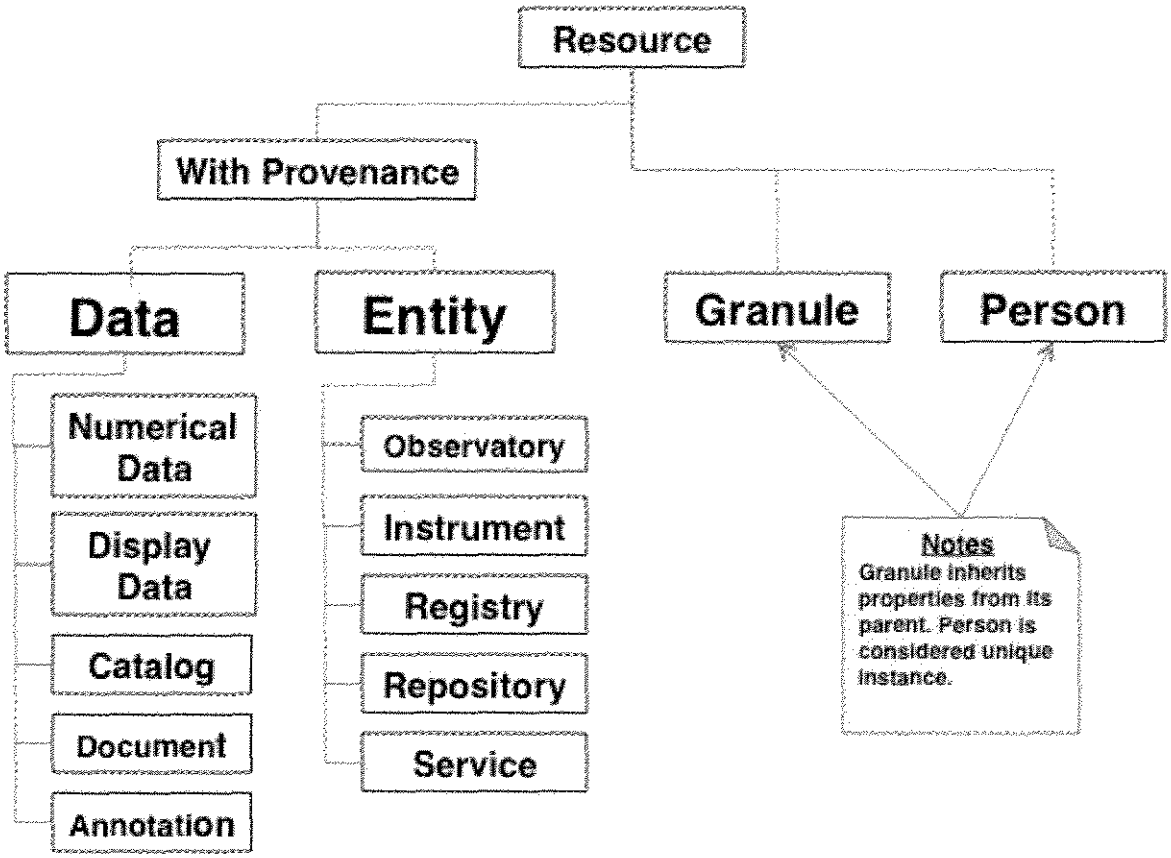

Figure 3. Information Categories, "Resources", and Subcategories within the SPASE Data Model

When many of the data sets are described according to the SPASE Data Model then one simple way to search for information throughout the data environment is to gather all of the descriptions into a single inventory system as indicated in Figure 4. The Virtual Space Physics Observatory (VSPO) is a system dedicated to this purpose. It contains all of the SPASE data descriptions that have been submitted from the various archives. VSPO can be kept up-to-date using a harvesting operation based on periodic searches for the latest SPASE data set descriptions. There are several methods for storage of the SPASE descriptions that have been written. The VMO, VHO, and VIRBO virtual observatories, for example, use "git" repositories for storage of their SPASE descriptions whereas the VITMO virtual observatory uses a relational data base management system for its storage. The various projects associated with VITMO provide the SPASE data descriptions while the XML-based data descriptions within VMO/VHO/VIRBO are supplied by "X-Men" who have experience with the SPASE Data Model and provide XML-based descriptions.

Figure 5 shows the VSPO interface for searching the entire collection of harvested SPASE descriptions. The searchable resources are shown toward the bottom of the left side column. There is also the possibility of searching for a time interval and/or a plain text string as indicated at the top of the left column. The results of the search are shown in the two rightmost columns and, if it is possible to download the data, a "Get Data" button on the right side appears.

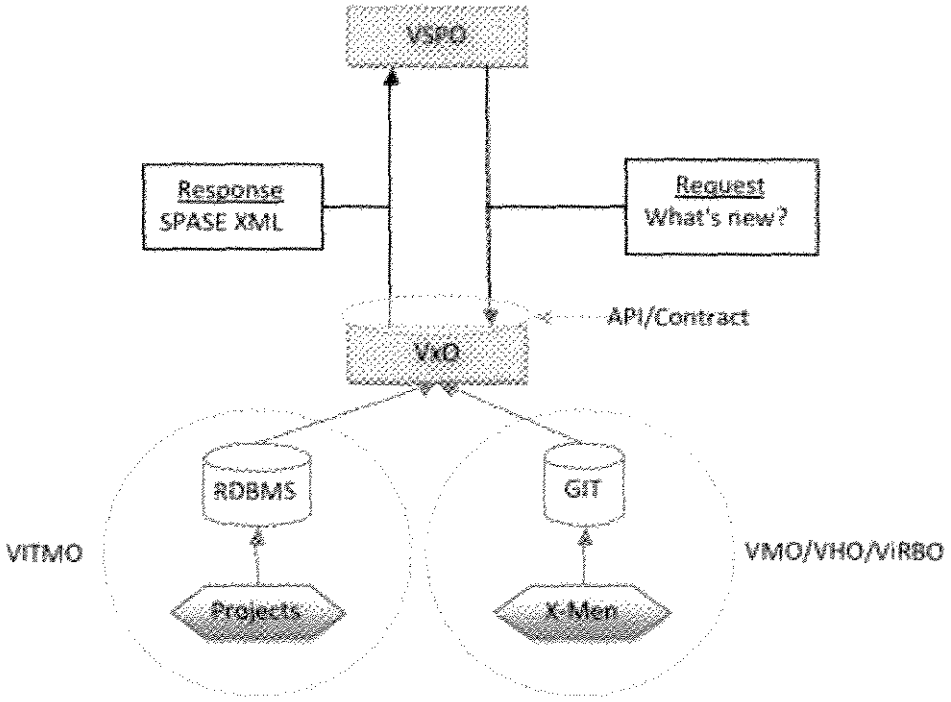

Figure 4. Harvesting Data Descriptions for inclusion in the Virtual Space Physics Observatory

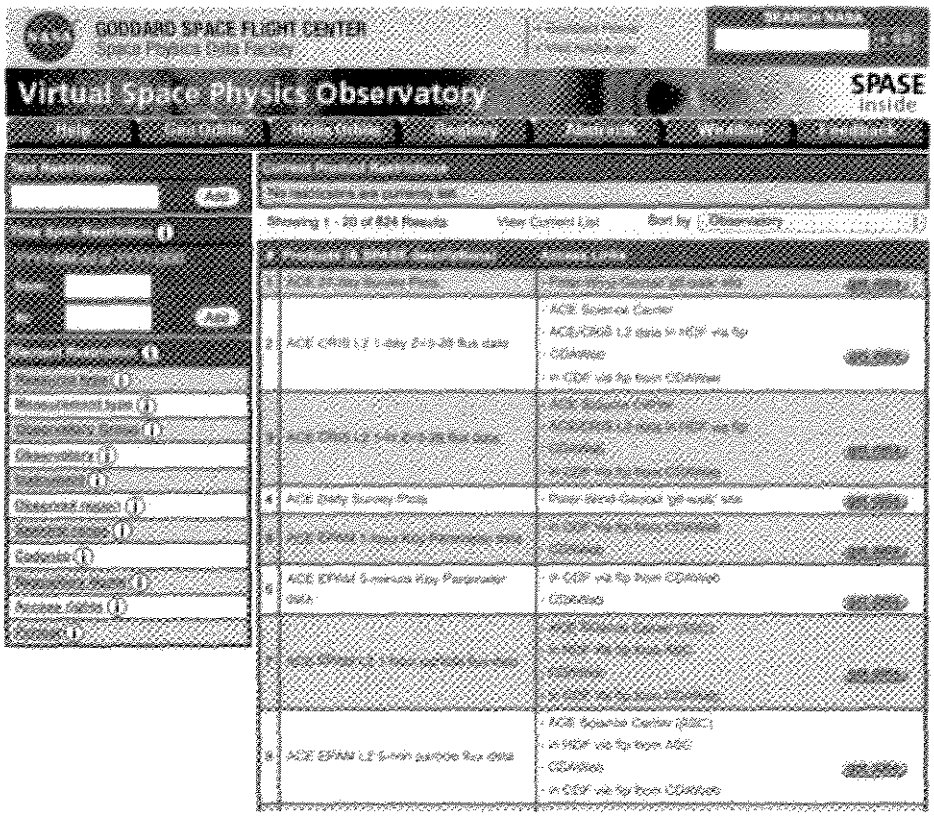

Figure 5. The Virtual Space Physics Observatory contains all SPASE data set descriptions and enables searching on SPASE resource metadata categories and subcategories.

Figure 6 illustrates the "Get Data" Accessor operation. A search request goes to the Virtual Observatory (VxO) and the registry associated with the Observatory is queried. The results are listed as a set of URL's that provide the links to where the data are available. The user invokes one or more of the links and a "Get" operation is sent to the specified data server for the archive of interest. The data are extracted from the associated repository and provided as a Response to the user. It is a simple yet effective method of data retrieval.

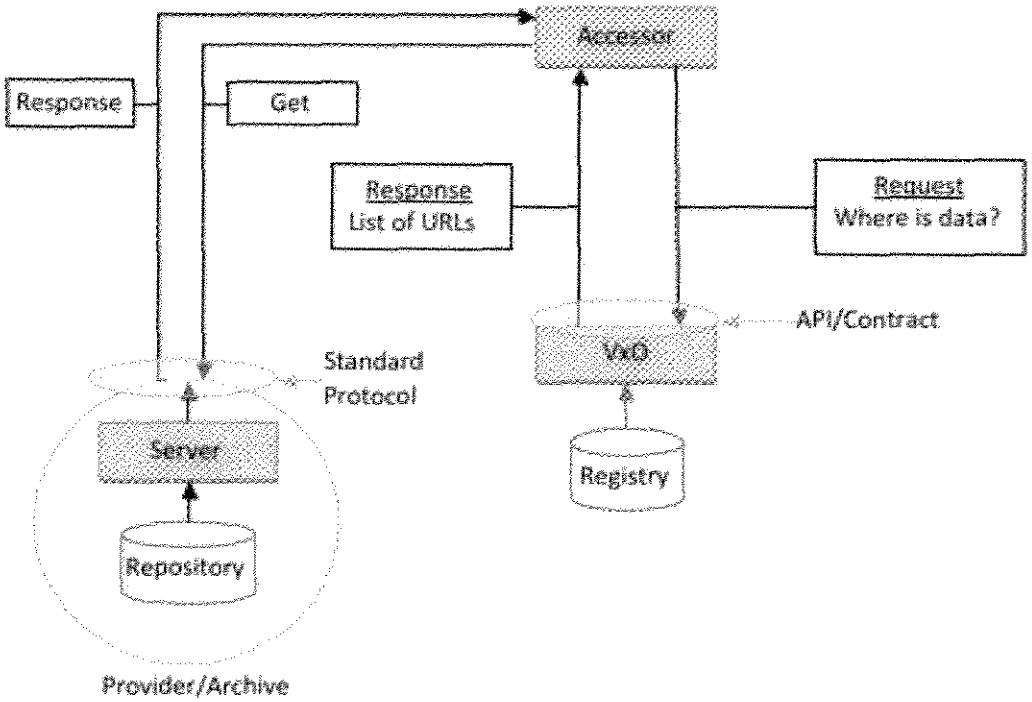

Figure 6. The data extraction process through SPASE queries to the Virtual Observatory (VxO)

Some would advocate that the simple retrieval of data is not enough. The data should be returned in a self-describing format sufficient to allow automated plotting and analysis of the data. This would require an extension to the Granule resource part of SPASE to support descriptions of the parameters contained in the data in a way that provides all the necessary auxiliary information for plotting and analysis. Even if all of this auxiliary information were made optional this would still complicate the SPASE Data Model and make it more difficult to comprehend for the first time user. Consequently the Granule description has been kept as simple as possible at this point in time. There is some capability of describing the individual characteristics of the parameters of the data and how they are laid out in the data record, but the feeling is that any additional complication of the Data Model will impede the progress of the use of SPASE.

SPASE PROGRESS

The main and most important question is whether SPASE is working as a catalyst layer to make the Heliophysics data environment interoperable. SPASE is currently in use and available to the solar and space physics community. The success of SPASE depends on widespread use within that community. That in turn is dependent on the willingness of the data holders to describe their data according to the SPASE Data Model. Unfortunately, it is human nature to avoid having to write descriptions of data. To ease the burden of writing the data descriptions the required aspects of a data description has been limited to as small an amount of information as possible. The simple, high-level descriptions are often sufficient for the purpose. Even those descriptions, however, require some time to write and are often subject to procrastination.

The NASA-supported Virtual Observatories are expected to provide SPASE format descriptions as one of the stipulations for the funding they receive. Recent surveys of the progress in SPASE description generation, however, indicate that the adherence to the stipulation varies widely among the Virtual Observatories. The Virtual Magnetospheric Observatory has a quite extensive list of descriptions generated and continues to generate them on a regular basis. Other facilities such as the Virtual Radiation Belt Observatory and the National Space Science Data Center have generated descriptions for some of their holdings, but still have a number yet to do. Still others, such as the Virtual Solar Observatory, have yet to start this task. The reasons behind this widely-varying response to the application of SPASE are many. In some cases the facility existed prior to the introduction of SPASE (such as the Virtual Solar Observatory) and is operating within its community quite well without the use of SPASE. It is not clear what advantage will be gained by translating existing data set information into SPASE.

So, what is necessary to advance the progress of SPASE? Acceptance of any standard within a community seems to occur once it is clear that the majority of the community supports or uses the standard. The mandate to use SPASE within the NASA-funded part of the Heliophysics community will help to achieve this perception, but there is a large portion of the community that does not receive NASA support, especially outside of the United States. It is that community that is of particular importance to convince to participate in SPASE. Some non-U.S. groups are already participating in SPASE such as the Centre de Donnees de la Physique des Plasmas (CDPP). The Cluster Active Archive, the Canadian Space Science Data Portal, the Inter-university Upper atmosphere Global Observation NETwork (IUGONET), various magnetometer chain groups, etc. There is continued effort to advertise SPASE and its value within the solar and space physics discipline so that others might know about and join the effort.

Another method to promote SPASE progress is to make it easier to write SPASE data set descriptions. Many tools have been developed for aiding the SPASE user in description creation and the use of the metadata descriptions. Example tools are:

- Ruleset Description Generator – create SPASE descriptions using external sources of information
- Web Editor – automated editor available through the network
- SPASE Assistant – a standalone editor for SPASE file checking
- Web+DB Editor – an editor that works with database storage
- XML Validate - determines compliance with a version of the SPASE data model
- Parser - convert SPASE XML to internal structures
- SPASE Registry Server - extracts information from SPASE registries
- SPASE Database Query - extracts information from SPASE resource descriptions
- Data Dictionary Lookup - converts or embeds SPASE metadata in other descriptions or forms
- SPASE-to-OAI mapping – embeds SPASE in Open Archives Initiative forms
- Correlator - divides an XML document into individual resource descriptions for a well organized file system
- Refcheck - determines the validity of all references in a resource descriptions; checks Resource IDs and URL

Other tools are still in development, such as:

- SPASE Query Language
- Java-to-XML Binding Language
- SPASE Guidelines Document

Many of the tools can be found by going to the website mentioned earlier (http://spase-group.org).

CONCLUSION

In summary, the space and solar physics community has a large number of small and diverse data holdings as well as a variety of larger data systems and services. The Virtual Observatories were

developed to provide some guidance for the community within a variety of subdisciplines of Heliophysics but it was the purpose of the SPASE working group to provide a common metadata language to unite the entire community similar to the way that the Flexible Image Transport System (FITS) unites the astronomy community. SPASE is in use now and Version 2.2.1 of the Data Model is currently available for download through the website. The document is of the order of 140 pages in length, but this includes many approaches to simplify communication of the basics of the structure of the model. It is understood that the key to progress in SPASE is in the widespread use of SPASE for data descriptions. A number of tools have been developed to simplify the process of creating SPASE data descriptions. The SPASE group is very open to new members and to modification of the Data Model to accommodate the needs of those who wish to use it. The main method of communication and change is the biweekly meetings via teleconference of the SPASE group. All are welcome to join these teleconferences to ask questions, express their opinions, make suggestions, etc. Contact the authors of this article for further information.