SPASE: THE CONNECTION AMONG SOLAR AND SPACE PHYSICS DATA CENTERS

J. R. Thieman[1], D. A. Roberts[2], and T. A. King[3]

*[*1] Code 690.1, NASA/GSFC, Greenbelt, MD, 20771 United States.*
Email: james.r.thieman@nasa.gov
[2] Code 672, NASA/GSFC, Greenbelt, MD, 20771 United States.
Email: aaron.roberts@nasa.gov
[3] IGPP, 5881 Slichter Hall, UCLA, Los Angeles, CA, United States.
Email: tking@igpp.ucla.edu

ABSTRACT

The Space Physics Archive Search and Extract (SPASE) project is an international collaboration among Heliophysics (solar and space physics) groups concerned with data acquisition and archiving. The SPASE group has simplified the search for data through the development of the SPASE Data Model as a common method to describe data sets in the archives. The data model is an XML-based schema and is now in operational use. The use is expanding, but there are still other groups who could benefit from adopting SPASE. We discuss the present state of SPASE usage and how we foresee development in the future.

Keywords: SPASE, Heliophysics, Archive, Data Model, XML Schema, Space Physics, Solar Physics

1 INTRODUCTION

The science of Heliophysics, otherwise known as solar and space physics, has been pursued using a variety of instruments both space-based and ground-based to gather data. Figure 1 indicates the many satellites that are now operational in space gathering these data. The picture does not show the many ground-based instruments such as magnetometers, radar facilities, ionosondes, etc. or many of the non-US satellites that also contribute to the accumulation of data. Within NASA the constellation of Heliophysics-related satellites is sometimes called the Heliophysics Great Observatory. With so many different instruments adding mountains of data to the data centers and archives around the world it is difficult for the researcher who may need data from multiple sources for a scientific study to find, retrieve, and analyze the data of interest. In some discipline areas the data repositories have been unified in data formats and methods have been put in place for finding data among all these repositories. Heliophysics has the problem of data being stored in many formats and in many repositories that are not part of a uniform discipline structure to facilitate data search and access. The Space Physics Archive Search and Extract (SPASE) project and the establishment of Heliophysics Virtual Observatories within NASA is an approach to solving this problem. Progress has been made and further progress depends on the acceptance and support by the non-NASA solar and space physics community in general.

Figure 2 gives some idea of the complexity of the Heliophysics Data Environment. Most of what is shown is the NASA-funded aspects of this Data Environment but there are some indications of the non-NASA parts which would complicate the diagram still more were they shown in full. There are many acronyms in Figure 2 and these are spelled out in Table 1.

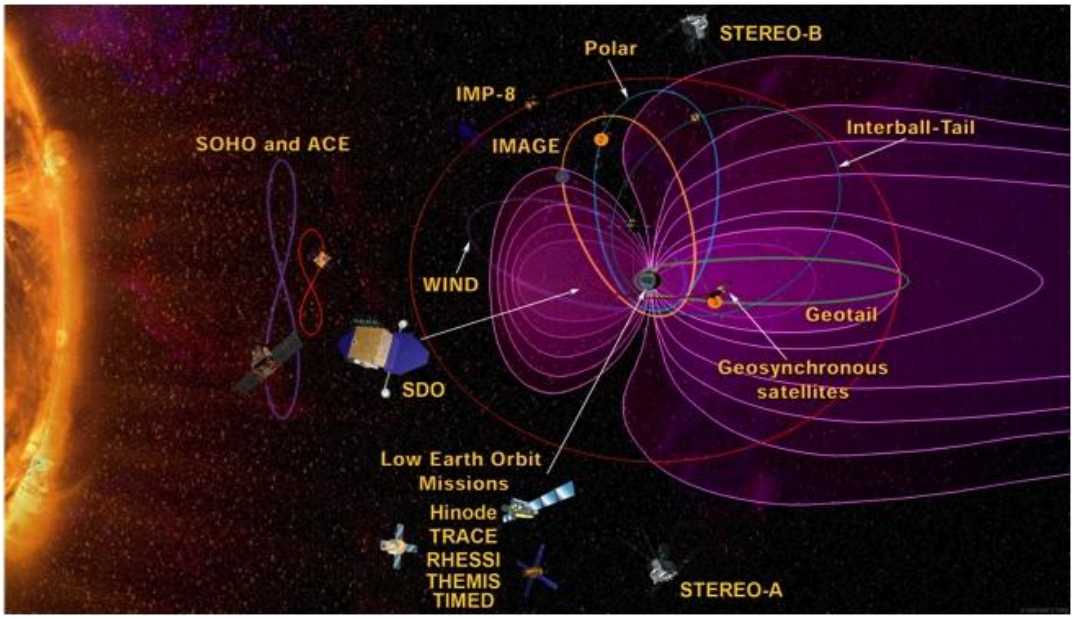

Figure 1. Heliophysics (Solar and Space Physics) Great Observatory

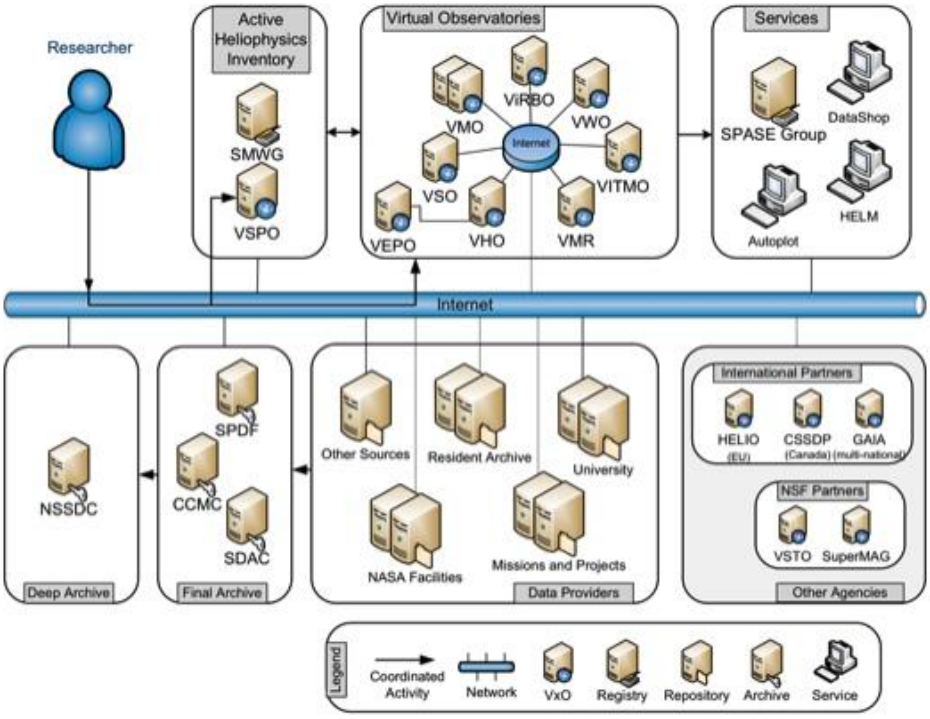

Figure 2. Heliophysics Data Environment

Table 1. Acronym list for Figure 2.

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

Figure 2 shows a researcher or user searching for data through the internet. Data are available through the NASA-funded data providers, other federal agency sources, university and other types of repositories as well as a variety of international partners as indicated toward the bottom of the diagram. Usually, each of these providers has different approaches to finding the data and a different user interface that has to be understood in order to locate what is needed. As the funding runs out for particular instruments the data are often transferred to a more comprehensive archive such as SPDF or SDAC and preservation copies are made and put into the Deep Archive at NSSDC. It is a daunting task to locate data of interest among all of these potential sources.

The NASA-funded Virtual Observatories (VxO's) were established to make this task easier within particular subdisciplines of Heliophysics. Using interfaces created by the Virtual Observatories the researcher can search for data within the subdiscipline and may be able to use special searches tuned to the needs of the subdiscipline users. The VxO's have the responsibility of knowing the data sources within their subdiscipline domain and providing a uniform approach to finding and acquiring the subdiscipline data of interest.

Problems arise for the researcher who wishes to compare or combine data of interest from several subdisciplines. Unfortunately, the VxO's do not have a common approach for access to data and the user is again faced with learning a variety of interfaces to do cross-disciplinary data access and retrieval. The lack of uniformity in finding Heliophysics data is partially associated with the variety of formats that are used to encode the data. There is not a single dominant data format in Heliophysics such as there is within Astrophysics in their use of the Flexible Image Transport System (FITS) format. A number of members of the international Heliophysics community agreed that it would be good to work toward a common metadata format and established the Space Physics Archive Search and Extract (SPASE) project to develop a uniform metadata approach across the discipline. With the Heliophysics data described using the common SPASE metadata format, metadata

inventories such as the SMWG and VSPO can be used to do searches for useful data across the entire discipline. This has been the main goal of the SPASE project

2 INTEROPERABILITY

Within the complex Heliophysics data environment what we ultimately would like to achieve is interoperability, making it easy for the user to search for and retrieve data and information. This has been the objective of many information systems through the years. The question is, how interoperable should we strive to make the various elements of the data environment? Levels of interoperability were proposed many years ago by the lead author and still may be of use in the discussion today. Figure 3 is an abstract diagram to facilitate the understanding of the levels of interoperability. The usual beginning situation is a group of very disparate elements of the overall data environment, each system element very different from the rest. The system elements are represented by the varying polygonal shapes in Figure 3.

Basic or Level 1 interoperability is achieved when these systems recognize that the user needs information or data from one of the other systems and provides a link for the user to follow to get to that system. Once the users reach the other system they are on their own in terms of learning the new system and how to find what they want.

Level 2 interoperability adds information to be passed to the new system so that the user's needs are at least partially known and the information can be used to help the user find what is wanted. Thus, the connections among the elements of the data environment become pipelines of information rather than simple electronic transfers. It is still necessary within level 2 to learn the nuances of the system to which one is transferred and this may not be straightforward even if the system has some information about the user needs to guide a search.

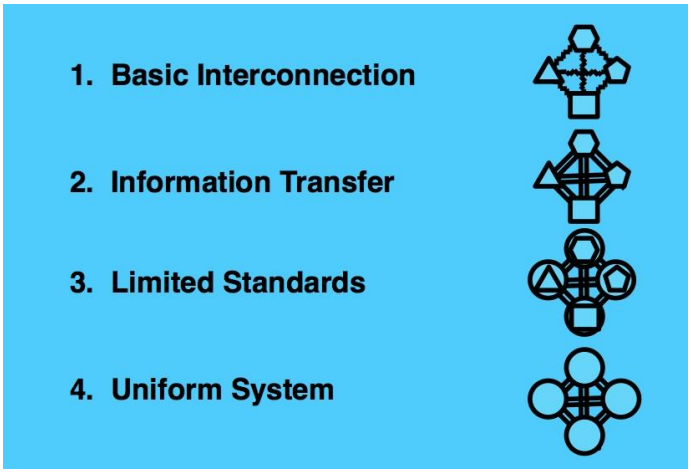

Figure 3. Levels of Interoperability

For Level 3 interoperability the systems are still very different from each other in essence, but they all agree to some standards of information exchange and/or the look and feel of the user interface. Thus the users encounter some familiar aspects among all the systems and are able to use previous knowledge to assist in an efficient search. There is a common "shell" that covers the disparate aspects of the systems. Thus, there is an added layer to each of the systems that provides some commonality among them.

Level 4 interoperability assumes that all of the systems agree to common standards and all of the systems are either created or modified so that to the user any differences among the systems are unnoticeable and it is as though the user were in one common system that has all of the needed information and data. The various elements of this environment may be physically located in very different places but that is irrelevant to the users since they are usually not aware that they have been utilizing different systems.

Levels 1 and 2 interoperability are relatively easy to achieve with internet capabilities as they are now and have been for many years. Level 3 interoperability is the more difficult step since it involves the process of getting many parts of the community to agree to common standards. Needless to say Level 4 interoperability is rarely achieved because of the need to tightly control all elements of the system. Level 4 may only be achievable if the

overall environment is built from the beginning according to exact standards. It is very expensive to modify existing systems to achieve this level of uniformity.

So, in the Heliophysics Data Environment the approach has been to try to achieve Level 3 interoperability through the adoption of the SPASE data model across all systems in the environment. The SPASE project is an international group of representatives of the Heliophysics Data Environment elements that has been developing the SPASE Data Model for many years. The SPASE Data Model is now available in Version 2.2.1 and the descriptive document as well as a variety of other information can be obtained from the main SPASE website at http://spase-group.org. This version of the SPASE Data Model has been stable for a relatively long time with only recent minor changes. Thus, the SPASE Data Model is in operational use especially among the VxO's within the NASA-funded Heliophysics Data Environment but still needs wider adoption among all the groups that are part of the global whole of this discipline.

3 THE SPASE DATA MODEL

The SPASE Data Model can be described as a grouping of Resources as indicated in Figure 4. The main Resources are those that describe the Data and the Entities associated with the Data. Most Data will be numerical in nature, but they could also be images or Display Data, Catalogs, Documents describing Data or just simple Annotations concerning the Data. The Entities associated with the Data are usually the Observatory (spacecraft, mission, project, etc.) that is the overall facility or group responsible for getting the Data and assuring the availability to the community and the Instrument that was used to acquire or generate the Data. Other Entities that may be associated with the Data are the Registry or Registries with information associated with the Data as well as the Repositories where the Data or copies of the Data are stored. Finally, Services can be associated with the Data which may be useful in interpretation or analysis. Both the Entities and the Data themselves will usually have Provenance information in order to track the origins of the Data and information attached to the Data. Another Resource of importance is the Person or Persons that were involved with the generation of the Data including the contact information needed to enable queries to the knowledgeable individuals.

The other Resource indicated in Figure 4 is the Granule. This is a subset of the overall set of data and usually represents a useful portion of the data for scientific analysis. Granules can be large or small and the number of Granules are often quite large. It becomes quite complex for the SPASE Data Model to describe the Granules in sufficient detail that the user has all the information necessary to analyze the data without having to ask the knowledgeable Person(s) for guidance. Some data formats, such as the Common Data Format (CDF), are self-describing in that they have the necessary information internal to the format to allow correct analysis. This self-describing property of a format is difficult to incorporate, however, and the question in the development of SPASE was whether this capability should be included when other formats are available that do this. This is a question that still is being discussed within the SPASE project. For the moment, SPASE has sufficient capability to describe the overall metadata associated with the Data, but does not yet have sufficient detail to fully describe and independently analyze the Granule subsets. This feature may not be included in SPASE since it would make the SPASE Data Model much more complicated and difficult to use for data description. It is argued that SPASE should be used for overall data finding and retrieval, but not for data analysis.

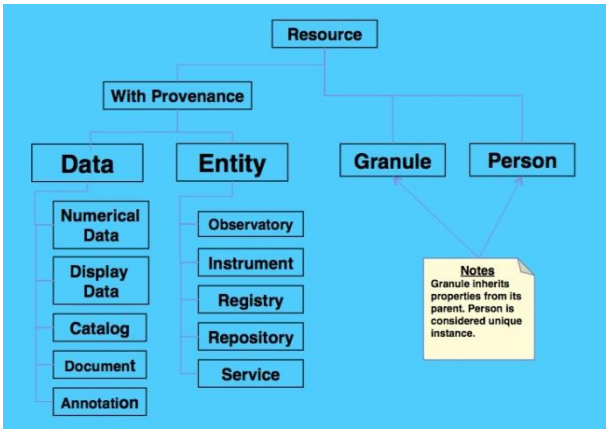

Figure 4. SPASE Information Model

4 HARVESTING AND EXTRACTION

Whether the SPASE data descriptions are used for data analysis or not, they can be gathered by any system through a harvesting process as indicated in Figure 5. The Virtual Space Physics Observatory (VSPO – http://vspo.gsfc.nasa.gov) in particular has the responsibility of maintaining a holding of all of the SPASE data set descriptions. Just as any system can do, VSPO periodically makes a request to the other VxO's for any new SPASE descriptions. These descriptions are usually stored in either a Relational Data Base Management System (RDBMS) as is done by VITMO or in a GIT repository as is done by VMO, VHO, and VIRBO.

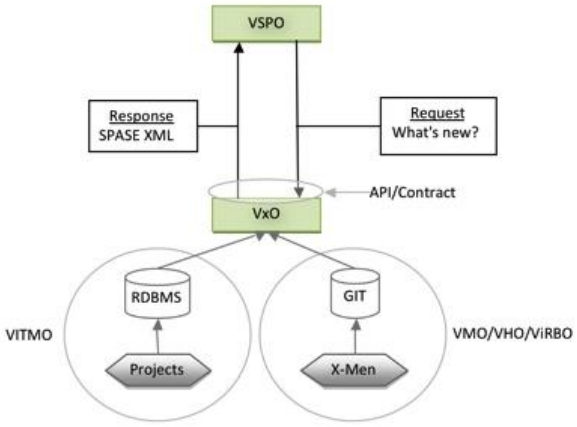

Figure 5. Harvesting of SPASE data descriptions from the VxO's and storage in the VSPO.

Generation of the SPASE descriptions is usually done by the projects or missions that have taken the data, but in some cases the VxO's have assigned personnel to do the generation of the descriptions. Since these descriptions are done in XML some of the personnel doing this work have taken on the title of "X-men". It is helpful to have experienced individuals creating the SPASE records, but certainly not necessary. Simple descriptions can be generated with just a basic knowledge of SPASE. The minimal SPASE record should contain enough information that a user can find it when searching for particular observatory or instrument names, personnel names associated with the data, and/or the generic parameters measured by the data.

When a user wishes to extract data of interest they may use any interface that is connected to a registry likely to contain the description for the data. If the query to the registry results in finding the description of the data of interest then there should be an "Accessor" (pointer or URL) that indicates an interest in acquiring the data. In the case of the VSPO this is a "Get Data" button that appears on nearly all data descriptions. When the Accessor is invoked by the user a query is sent to the server connected to the data repository and a response is sent to the user either containing the requested data or indicating how the data may be accessed. Figure 6 is a diagram of this data extraction process.

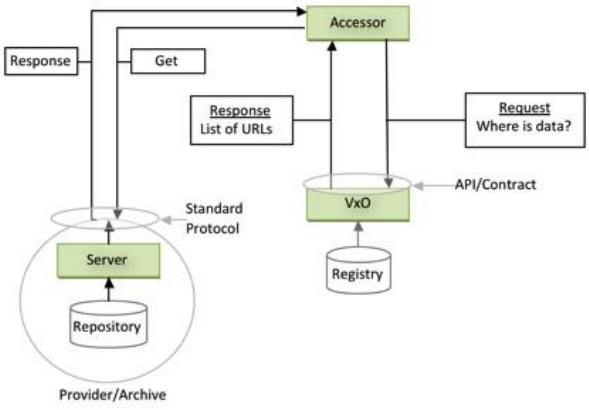

Figure 6. Extraction of data from the VxO's via the SPASE data description.

The success of SPASE depends, of course, on the willingness of the data holders to describe or have their data described using the SPASE Data Model. However, it is usually human nature to not wish to spend time writing descriptions of data holdings, even if it is just a simple high level data description. For this reason a number of tools have been created to help with the data description process. These include: a Generator which can create SPASE descriptions based on Rulesets and external sources of information; several types of Editors which work through the Web or in standalone form, or Editors that use database storage; and a Validator which will check compliance of a SPASE description with the latest version (or earlier versions) of the SPASE data model. Many other tools exist as well and are generally available from the SPASE website (http://www.spase-group.org).

If SPASE is brought into widespread usage then it will bring about a relatively uniform approach to the data access within space and solar physics. The data systems information will be accessible in a relatively uniform approach making the data more easily accessed and reaching a level 3 type of interoperability as discussed earlier. Can level 4 be achieved where it seems as though all of the systems are uniform and appear to be part of one common system? Perhaps this might be attained through the widespread use of cloud computing technology which is gaining popularity rapidly. Could a common cloud storage of data and information provide uniformity to the Heliophysics discipline? It is not clear if this is achievable at the present time or even desirable, but it is something to keep in mind. Whether level 4 is reached or not, the general application of SPASE provides a needed connection among the disparate parts of the solar and space physics data community.

5 CONCLUSION

To summarize the contributions of the SPASE project and the Data Model created by the project we review several points. The Heliophysics data environment has historically been very diverse and globally dispersed, but it is now being unified through advancements such as that provided by the SPASE Data Model through a standard metadata approach. The key to this approach is the creation of data descriptions in accordance with the Data Model. It is not easy to get the Heliophysics community to create these data descriptions, but the use of the SPASE tools which have been created for facilitating this process can greatly influence the progress toward creating the descriptions for all data of interest. If this were achieved it would provide a common link within the data environment that would be a type of level 3 interoperability as defined earlier. With the commonality provided by the SPASE Data Model cross-disciplinary research becomes easier and this is a step toward still more uniformity. Will a completely uniform level 4 type of interoperability ultimately be achieved through cloud computing and/or other technologies? This would be an interesting step, but may not be worth the extra effort necessary to achieve it. The main emphasis in the immediate future is the further adoption of the SPASE approach by the global space and solar physics community. Interested readers are invited to contact the authors for additional information.

6 ACKNOWLEDGEMENTS

The lead author would like to acknowledge the contributions of his co-authors in their suggestions for the presentation and especially for the use of some of the graphics they had generated.