The Space Physics Archive Search and Extract (SPASE) and the Virtual Space Physics Observatory (VSPO)

James Thieman[1], Todd King[2], Aaron Roberts[3], Joseph King[4], Christopher Harvey[5], Christopher Perry[6] and Joey Mukherjee[7]

[1] Code 690.1, NASA/GSFC, Greenbelt, MD 20771, USA
james.r.thieman@nasa.gov
[2] Institute of Geophysics and Plasma Physics, UCLA, 5881 Slichter Hall, Los Angeles, CA 90095-1567. USA
tking@igpp.ucla.edu
[3] Code 612.2, NASA/GSFC, Greenbelt, MD 20771, USA
roberts@vayu.gsfc.nasa.gov
[4] QSS, Inc., Code 690.1, NASA/GSFC, Greenbelt, MD 20771, USA
jking@mail630.gsfc.nasa.gov
[5] CESR-CNRS, B.P. 4346, 9, avenue du Colonel Roche
31028 Toulouse CEDEX 4 – France
harvey@cesr.fr
[6] Rutherford Appleton Laboratory, Chilton, OXON, OX11 0QX, United Kingdom
c.h.perry@rl.ac.uk
[7] Southwest Research Institute, 6220 Culebra Road, P.O. Drawer 28510, San Antonio, TX 78228-0510
jmukherjee@swri.org

Abstract. The Space Physics Archive Search and Extract (SPASE) System and the Virtual Space Physics Observatory (VSPO) are NASA-funded, collaborative development efforts among multi-institution, international, space physics data holding organizations. They are intended to allow science users to interoperably locate space physics data of interest, intercompare the information about data, and retrieve selected data sets or portions of data sets. The SPASE data model provides a common metadata language to be used in searching the distributed archives. VSPO uses the data model to describe the data resources in the space physics community and provides an interface to find and access these data.

I Introduction

The SPASE Data Model (see http://www.igpp.ucla.edu/spase/) has been developed to provide a common metadata language for searching across the multitude of disparate space physics data archives for data of interest. The Virtual Space Physics Observatory (VSPO – see http://vspo.gsfc.nasa.gov/websearch/html/VSPO.html) is an

example implementation of the SPASE data model describing important data holdings of space physics data archives in many places. VSPO users can generate a query based on the common SPASE metadata language. Presently, the query searches the VSPO database of SPASE-compliant data descriptions. The query provides a list of data sets from which the user may choose one or more of interest and either follow links and local data archive procedures to acquire the data or, for a significant number of VSPO-identified data sets, have VSPO retrieve and deliver the desired subset of data. The group is working to make the data set description process more automated for data providers. SPASE is intended to facilitate intercomparison of search results by displaying results using common time or location formats. Translation software underlies a uniform interface, insulating the user from the varied search procedures at the individual institutions. The SPASE model supports data descriptions at multiple levels of detail allowing providers to describe their holdings in a way that supports access of selected parts of data sets if enabled for the user. We are working with other Virtual Observatories and data centers relevant to space physics to enable data searching, access, and retrieval across multiple disciplines. Development of an effective data model and data retrieval process is challenging and requires feedback from all sectors of the community. We welcome new participants and feedback from the scientific and technical data community on the best approaches to assuring preservation and continued usage of our data repositories into the future.

Increasingly, studies of the Sun, the Earth, and their extended plasma and field environments involve multiple datasets from many spacecraft, ground-based observatories, and models. These science products and the software and documentation associated with them are stored in a wide range formats at diverse repositories around the world. A scientist wishing to find out about products outside of his or her immediate expertise will typically find it both difficult to locate data and to interpret the data once found. On the other hand, the diverse products can usually be described in physical terms that are part of the common language of science. The data model provides a set of terms that, ideally, can be mapped onto the specific terms used for and in products, thus providing a uniform means of access and description. The data model is intended to provide the cornerstone of "Virtual Observatories" that will link broad ranges of solar and space physics in a natural way.

More specifically, the data model should:

(1) Provide a way of registering products using a standard set of terms that allows the products to be found with simple searches;
(2) Allow searching for products containing particular physical quantities (e.g., magnetic field; spectral irradiance) that are represented in a diverse way among the data products; and
(3) Create a means of mapping comparable variables from many products onto a common set of terms so that visualization and analysis tools can be used on all of them without regard to the origin of the data.

The accomplishment of these tasks requires "middleware" that understands product registries and performs the translations needed to map the idiosyncratic file names and quantities of each repository onto standard terms. This intermediate layer, which can

take many forms for different purposes, will provide the links necessary to connect user applications and search-and-retrieval front ends to data repositories. Ultimately, the data environment centered on the data model will involve a number of software tools as well, linked together as internet-based services or other means. Specific software tools and documentation associated with products will be accessible straightforwardly. This "system" has the potential to provide capabilities that can aid even expert users of a particular dataset (e.g., on-the-fly coordinate transformations or the ability to merge datasets from different instruments), in addition to providing the broad access needed to solve global problems in Sun-Earth connection physics. Success will require a concerted cooperative effort across disciplines. Existing efforts in Space and Solar Physics as well as in other areas such as Earth Science and Astronomy will guide the work.

2 Development of the Current SPASE Data Model

The data model presented here grew out of the efforts begun a number of years ago that became formalized in regular teleconferences of a group of interested data providers, including scientific and technical representatives of some of the largest space physics data holdings in the US and Europe. The participating institutions presently involved in the effort are:

- NASA/Goddard Space Flight Center – National Space Science Data Center (NSSDC)
- Centre de Donnees de la Physique des Plasmas (CDPP) in France
- UCLA Institute of Geophysics and Planetary Physics (IGPP)
- Rutherford Appleton Laboratory (RAL) in England
- Southwest Research Institute (SwRI)
- Applied Physics Laboratory (APL)
- Jet Propulsion Laboratory (JPL)
- Augsburg College
- NOAA/National Geophysical Data Center (NGDC)
- European Grid of Solar Observations (EGSO)
- Institute of Space and Astronautical Science (ISAS) in Japan

As the effort to provide seamless access to distributed data proceeded, it became clear that the data model efforts were central. Thus, in March of 2003 a meeting of many of the participating institutions was convened to begin the data model construction in earnest. The initial effort involved collecting terms from CDPP, SwRI, NSSDC and other sets to form a starting point. A year of teleconferences, e-mailed revisions, and occasional face-to-face efforts, along with the application of the terms to specific cases, has led to the present document.

The general philosophy of the data model is to describe products using a natural taxonomy of data sources and of the physical world as represented in actual or potential datasets and models. The resulting data model has been put to many tests, but will have to evolve as new products are considered.

3 The Data Model in a Virtual Observatory Context

The current conceptual model is intended to serve as the basis for interoperability between independent data systems. For a .PDF copy of the data model document click on "Current Draft" on the SPASE home page at http://www.igpp.ucla.edu/spase/. The model is an attempt to capture the various concepts that are used to represent the shared knowledge in the Space and Solar Physics domain. The concepts are used to convey queries and responses between the user and various data systems. Many such systems, in diverse fields, are now functioning or planned. The paradigm for these is the "Virtual Observatory" (VO) originally suggested by the astronomical community that would allow any researcher anywhere in the world to access all known observations of a chosen patch of the sky using the Internet to query repositories of information distributed around the world at data centers and observatories.

Usually, the community of a particular science discipline comes together to form or support a Virtual Observatory. Through a VO users can gain access to a wide variety of data in archives throughout the discipline community. Each VO is tailored to the needs of that particular community. The VO includes one or more interfaces that facilitate data access, especially for researchers in the community. There can be multiple VO's in a particular community in which case it will be helpful to come to a common underlying scheme for data search and access. In the space physics community it is intended that the SPASE data model provide the underlying commonality. The SPASE group is working closely with the Virtual Space Physics Observatory (VSPO) to test and implement the data model. The SPASE group is also working with the Virtual Solar Observatory (VSO) and Virtual Solar Terrestrial Observatory (VSTO) with the intent of adopting SPASE as a common data model. There are Virtual Observatory efforts going on in many science disciplines now and the overall group of these are sometimes referred to as the "VxO's".

The concepts presented in the data model are independent of any implementation, but we have in mind a likely map of the eventual data environment. Figure 1 illustrates a possible architecture; the details will depend on the outcome of a number of current efforts and their coordination.

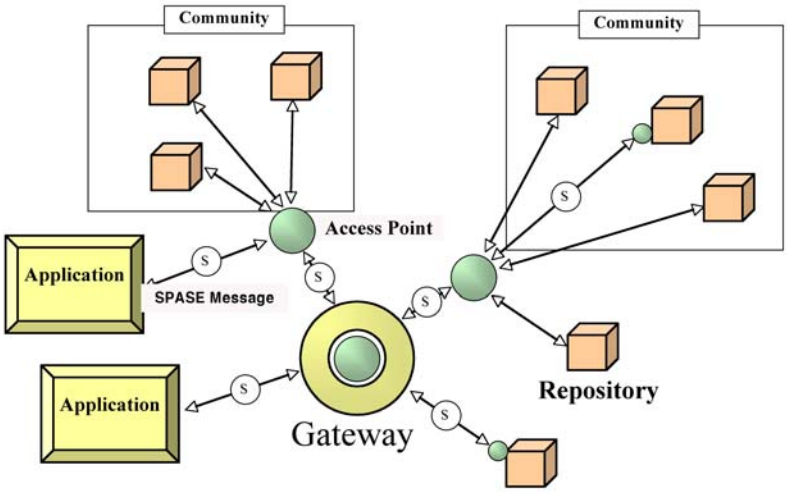

Fig. 1. A possible data environment architecture. Information and data flows from repositories to applications through access points and gateways. The portions of the system using SPASE-Data-Model-based messages are indicated with the Ⓢ.

A prototypical usage scenario is:

1. The user defines a query using an application.
2. The application contacts relevant participating access points or a general "gateway".
3. Each access point or gateway queries its database and generates a reply.
4. The application uses the reply to either present metadata to the user (a URL or a request for further definition of the query) or to stream data back to the application.
5. After further refinement of the query, the user can request files or subsets of files to be returned to the application or to the user directly.

We envision that different control authorities will maintain different aspects of the data model. For example, the name of observatories is maintained by NSSDC; the definition of prime meridian used in describing local latitude and longitude is set by IAU; and the accepted format for time representation is defined by ISO. There will need to be a community approved group to be the central authority on the data model; SPASE is currently serving as a prototype for this role.

It remains to be determined the extent to which registries of products will be centralized such that one location will contain the latest list obtained from registered data providers. There should at least be a common format, if only keyword=value lists, for the exchange of product registries. XML schema may work if they are widely adopted and understood.

4 The Virtual Space Physics Observatory

Current plans at both the NASA and the community-wide Sun-Earth connection level include making data easily available from all missions relevant to the global problem of the effects of solar particles and fields on the Earth. The VSPO is an evolving system for accomplishing this task. The basic philosophy, shared with the Virtual Solar Observatory (VSO) and many other such projects, is to register data products from disparate repositories using a common language that allows searching across datasets in a uniform way. This can be done in several steps.

First Step: A Space and Solar Physics Product Site Finder

For the first instance of a VSPO, a Space Physics Product Site Finder has been developed (http://vspo.gsfc.nasa.gov/websearch/dispatcher) - a website that allows users to quickly find data files (e.g., on an ftp site) and interfaces to data from a wide variety of missions, as shown in Fig. 2. The goal of the Site Finder is to provide a direct link to data and mission/instrument sites. The Site Finder interface enables the user to formulate complex queries and quickly find the desired data products. The user has an option to construct queries using a specific set of dictionary terms, or by using the text search feature. The Site Finder has been initially populated using publicly available information on many products. Presently the data descriptions are not as complete as they should be. Suggested changes to these descriptions should be sent to co-author Aaron Roberts by e-mail.

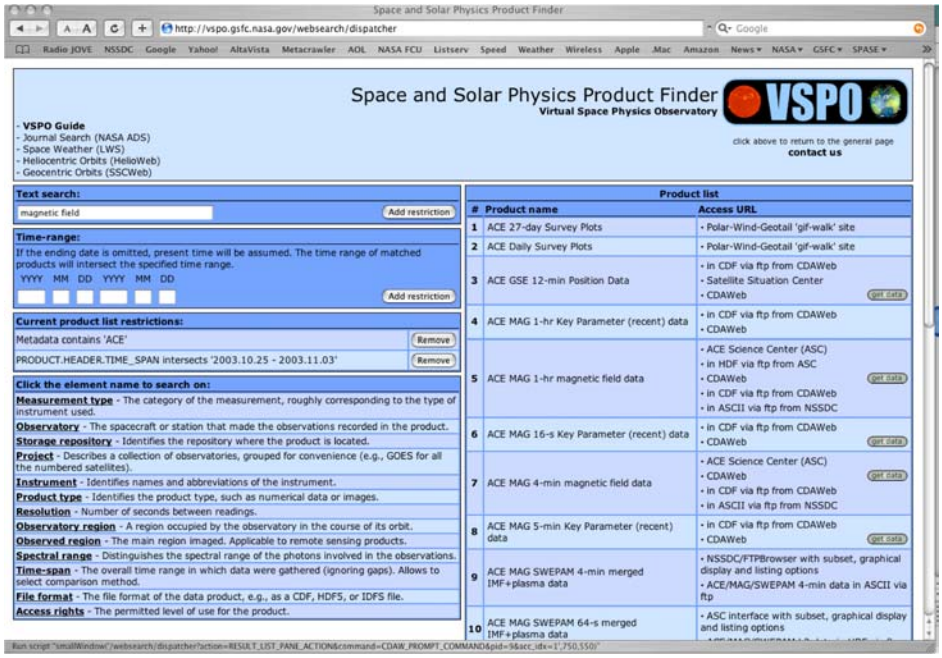

Fig. 2. The VSPO interface for finding data resources among widely distributed data centers using a variety of search criteria.

Second Step: Direct Use of Data Services

We have begun extending the Site Finder to include the ability to use the data services of various repositories directly. The Coordinated Data Analysis Website (CDAWeb) now has a Simple Object Access Protocol (SOAP) Web Services interface that we are using to allow users to obtain Common Data Format (CDF) files, ASCII files, and graphs of selected data subsets using the services provided by CDAWeb. For some purposes (such as simultaneous multiple product retrieval), CDAWeb offers more capabilities, but the VSPO service should be convenient for quick access without changing to another interface. There is also an intent to provide access to VSO, the ACE Science Center, and other VxOs and repositories now that the first case is working. The VSPO will continue to be developed and enhance toward a proposed Virtual Sun Solar System Connection Great Observatory (VS3O) architecture as shown in Fig. 3.

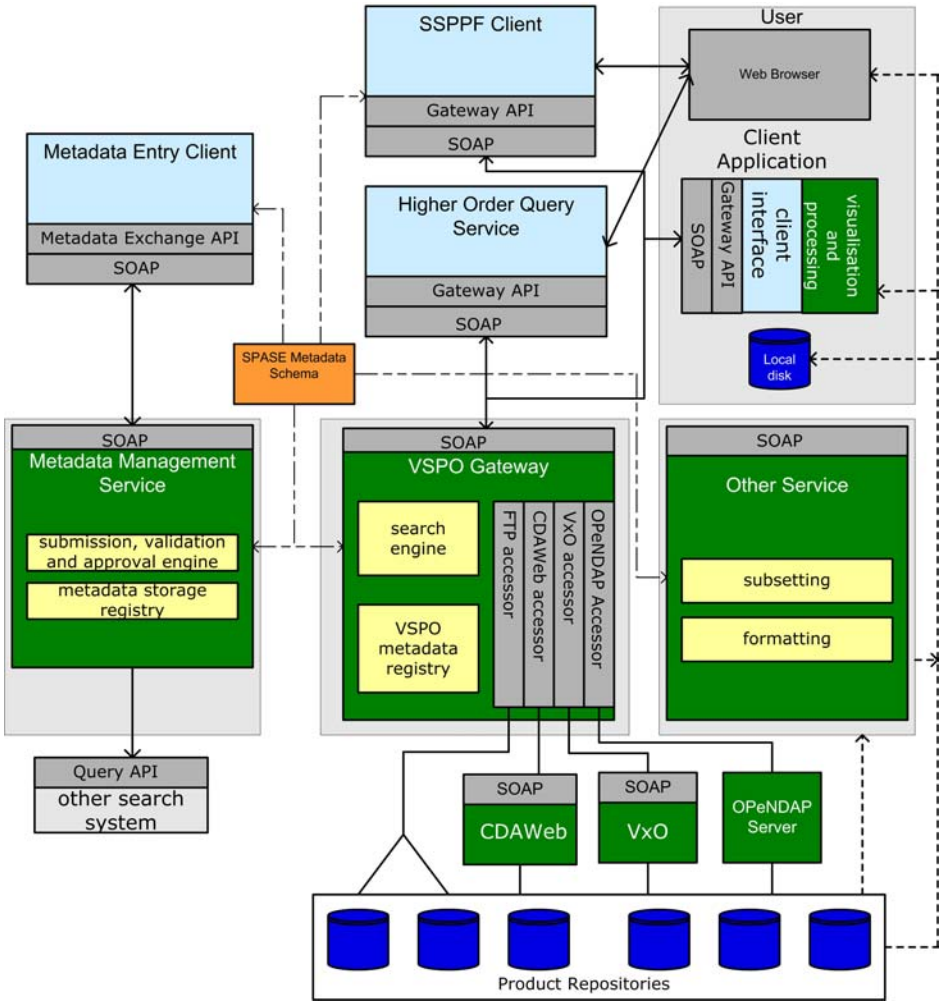

Fig. 3. VS3O system architecture diagram showing extended capability for the VSPO system.

The next step for VSPO will be to provide more direct links to the repositories using ftp and more flexible "SOAP" interfaces, as in the Second Step above. Farther down the road, the functionality will be extended to include requesting and receiving subsets of the data using VSPO-based services. The longer-term plan is to provide access to additional services such as coordinate system transformation, format translation, and use of event lists in a "higher-order query" service. All this work is being carried out in collaboration with many members of the space physics community, and it will only be through having input and suggestions from a wide variety of people that this and many related efforts will succeed. The Living With a Star Data Environment page (http://lwsde.gsfc.nasa.gov/) provides much more information on community efforts.

5 Summary

SPASE and VSPO are important steps to providing ready access to data for the space physics community, but they depend on community feedback and acceptance for the broad application that will make them valuable. All are welcome to download the data model document and try out the VSPO interface as described earlier. Comments are welcomed by the authors. Feel free to contact any member of the co-authors if you have further interest.