SPASE 2.0: Standardization of Space Physics Data Access and Retrieval

Jim Thieman[1], Todd King[2], Aaron Roberts[3], et al.

(1) NASA Goddard Space Flight Center

Code 690.1, Greenbelt, MD, United States

EMail: James.R.Thieman@nasa.gov

(2) Institute of Geophysics and Planetary Physics

University of California, Los Angeles, CA, United States

EMail: (Tking@igpp.ucla.edu)

3 NASA Goddard Space Flight Center

Code 672, Greenbelt, MD, United States

EMail: Aaron.Roberts@nasa.gov

ABSTRACT

SPASE stands for Space Physics Archive Search and Extract. Over a number of years the international SPASE Working Group (see http://spase-group.org) has developed the SPASE Data Model which is a standard for describing data sets in the Space and Solar Physics (now often called Heliophysics) domain. The Data Model is a set of terms and values along with the relationships between them that enable describing all the resources in a heliophysics data environment. It came about through an effort to unify and improve existing Space and Solar Physics data models. The intent of this model is to provide the means to describe resources, most importantly scientifically useful data products, in a uniform way so they may be easily registered, found, accessed, and used. The main method of data discovery is through international Virtual Observatories (VO's). SPASE released version 2.0 of its Data Model in April 2009. It has been used to document a large range of data sets of interest to the space physics community. Version 2.0 of the SPASE Data Model provides a solid foundation for future development. It contains many inherent capabilities which are yet to be realized. Potential applications are discussed to illustrate the range of capabilities.

Keywords: Heliophysics, data, model, archive, interoperability, metadata, resource, description

INTRODUCTION

The Space Physics Archive Search and Extract (SPASE) consortium is an organization to set community-based standards with the goals of defining a data model for Space Physics (now often referred to as Heliophysics), demonstrating its viability, and enabling interoperability in a federated data environment. The main purpose of the work is to enable space physics data resources to be easily registered, found, accessed and used.

The recognition that a coordinated effort for the space physics data environment was needed was first discussed at the International Solar-Terrestrial Physics meeting at the Rutherford Appleton Laboratory in the United Kingdom in 1998. Informal, grassroots efforts were carried on within the community for the next seven years to come to a unified approach for data archiving and retrieval. It was also recognized that a more intense effort with personnel spending a significant fraction of their time on this effort was needed and, in 2005, a proposal to create the Spase Physics Archive Search and Extract project was funded through the NASA Living With a Star program.

The group that made up the core of the project quickly determined that the main part of the SPASE effort should be devoted to a common SPASE Data Model to serve as a metadata medium of information and data exchange among the widespread and diverse data archives within the community.

Although the core group that received funding through the SPASE proposal was not large, the number of people who have been involved in the development of SPASE up to the present is more than 50 and the number of institutions that they represent is more than 15.

The first operational version of the SPASE Data Model (version 1.0) was released in 2005. In 2006 NASA solicited proposals to establish thematic virtual observatories for the heliophysics community and SPASE was adopted as the metadata standard to enable interoperability. In the following years, through more than 4000 emails, greater than a hundred biweekly teleconferences, and at least six face-to-face meetings, the SPASE Data Model continued to be developed and the latest version is now 2.0.2 as of September, 2009.

VERSION 2.0 OF THE SPASE DATA MODEL

The SPASE data model is presently being used to describe many data sets within the heliophysics domain. In particular, the Virtual Observatories' staff are creating data set descriptions as an indication of their data holdings and to enable the worldwide space and solar physics community to find and retrieve data of interest. Thus, researchers in the magnetospheres, waves, ionosphere-thermosphere-mesosphere, radiation belts, energetic particles, solar physics, and even the modeling and simulation communities are using the SPASE Data Model to describe data and providing feedback on their experiences. The biweekly teleconferences of community representatives continue to treat problems and modify the data model as concerns are expressed.

Also, SPASE-related services are being developed that improve the utility of the model for the community. These include the development of resource registries which provide readily-available information on commonly-needed resources such as spacecraft or observatories, instruments on the spacecraft or observatories, personnel associated with important data sets, etc. A distributed query language based on SPASE metadata conventions, called SPASE-QL, is being developed to allow more versatility in the use of SPASE data descriptions for the search and retrieval phases important to the general user.

Figure 1 shows a conceptual model of the concepts and resources contained within the SPASE Data Model. Although this is not an ontology it gives an overview framework for the Model.

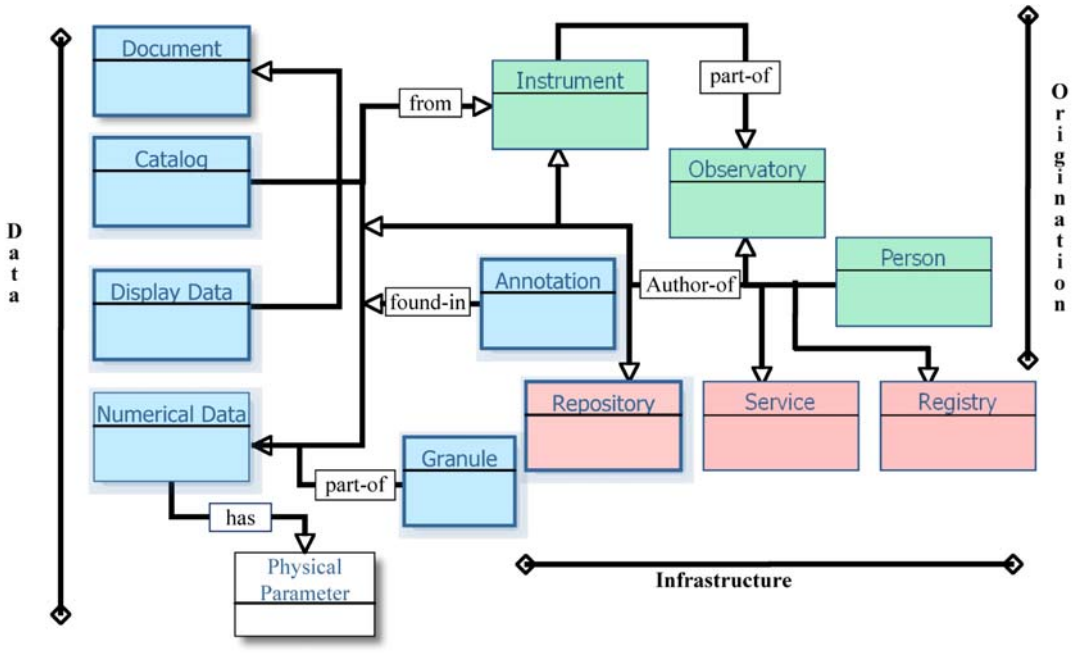

Figure 1. A simplified conceptual model of resources (classes) in the SPASE data model. Arrows point in the direction of association. Cardinality is not shown.

The SPASE Data Model was developed by a team of scientists, information technology specialists, data engineers and software developers. The development group is mainly composed of users or those who are very closely associated with users, so it is not surprising that the documentation is oriented toward the known users and the terminology they use on a regular basis. Consequently the nomenclature of the Model is less formal than the standard Unified Modeling Language (UML). So, for example, the SPASE Data Model term "Resource" maps to the term "Class" in UML. Similarly "Element" in SPASE would be mapped to "Attribute" and "Container" to "Component Class".

Despite using what might be considered non-standard terms, SPASE does adhere to the principals of the ISO 11179 Metadata Registry standard without formally adopting all aspects of the standard. So, in accordance with part 1 of the ISO standard, SPASE has a Data Element Concept, a Representation, and Data Element Relationships. SPASE maintains a simple metadata registry which is in line with part 3 of the ISO standard. Likewise the Data Model has Data Element Formulation Rules and Naming and Identification Principles as noted in parts 4 and 5 of the standard respectively.

As work on the Data Model continues there are certain design principals to which the SPASE consortium tries to adhere. These are:

- **Data are self-documented -**
  Data resources have internal schema or structures for storing values;
- **Resources are distributed -**
  There are many providers of resources and these providers can be located anywhere in the world.
- **Online Resources have Universal Resource Locators (URL's) -**
  If a resource is on-line it can be accessed and retrieved using Universal Resource Locators (URL); and
- **The data environment is continuously evolving -**
  New resources are actively generated either as part of an on-going experiment or as a result of analysis and assessment.

The SPASE Data Model is a semantic data model in that it defines the meaning of data within the context of the interrelationships with other data. The Model defines the scientific context of the data. Although there is not a completely defined ontology, there are typed associations between the resources. The Data Model does not depend on any particular implementation, although the usual implementation that is used for data description follows an XML schema. There are numerous XML stylesheets available for converting the metadata expressed in XML to other forms such as Hyper Text Markup Language (HTML) or the protocol advocated by the Open Archives Initiative (OAI).

VIRTUAL OBSERVATORIES AND THE HELIOPHYSICS DATA ENVIRONMENT

The worldwide heliophysics data environment is complicated by a large number of data archives and many satellite and ground-based data sets both large and small. To make these data more readily available and usable a "virtual observatory" (VO) system has been developed within the heliophysics community. This system works together with the existing and newly-created archives as well as data providers and services that have been established within the community. Figure 2 provides a simplified overview of the data environment, all tied together by the internet. There are a large number of acronyms associated with this diagram, so these are spelled out as indicated below.

Active Heliophysics Inventory

- SMWG – SPASE Metadata Working Group
- VSPO – Virtual Space Physics Observatory

NASA-Funded Virtual Observatories

- VEPO – Virtual Energetic Particle Observatory

- VHO – Virtual Heliospheric Observatory
- ViRBO – Virtual Radiation Belt Observatory
- VITMO – Virtual Ionosphere, Thermosphere, Mesosphere Observatory
- VMO – Virtual Magnetospheric Observatory
- VSO – Virtual Solar Observatory
- VMR – Virtual Model Repository
- VWR - Virtual Wave Observatory

Services

- HELM – Heliophysics Event List Manager

International Partners

- CSSDP – Canadian Space Science Data Portal
- GAIA – Global Auroral Imaging Access
- HELIO – The Heliophysics Integrated Observatory
- VSTO – Virtual Solar Terrestrial Observatory

Final Archive

- CCMC – Community Coordinated Modeling Center
- SDAC – Solar Data Analysis Center
- SPDF – Space Physics Data Facility

Deep Archive

- NSSDC – National Space Science Data Center

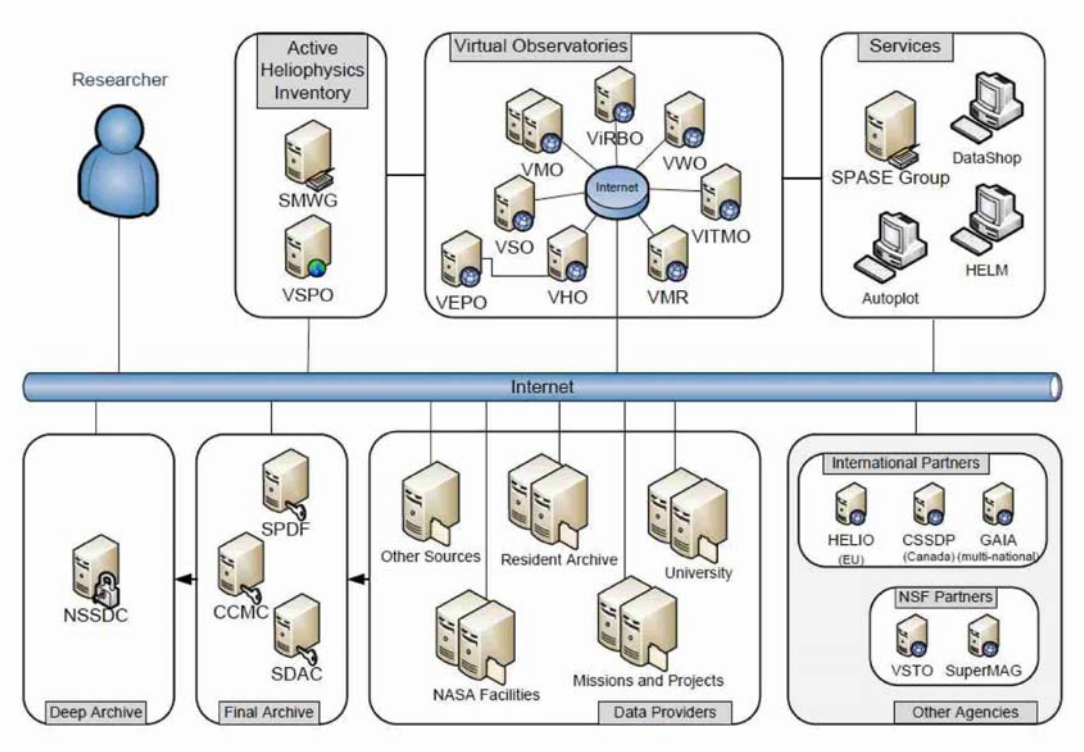

Figure 2. A simplified diagram of the Heliophysics Data Environment for data location, archiving, and retrieval.

FACILITATING DATA MODEL APPLICATIONS

The main task for the SPASE group at the present time is facilitating data descriptions for the important data sets within the heliophysics community. This is ongoing, especially by the staff of the various NASA-funded virtual observatories. The SuperMAG project is associated with the National Science Foundation and is represented as well. As indicated in Figure 2, there are also non-US participants in Canada and in the European Union. To make the task easier and the results more useful to everyone a number of tools have been developed to help create the data descriptions and make them more useful to the community once they are in existence. In most cases there tools are meant to work with the reference implementation of the model in XML. The types of tools include:

- Validator to determine compliance with a given version of the SPASE Data Model;
- Parser to convert SPASE XML to other internal structures;
- Editor to aid the creation of SPASE descriptions by hand;
- Generator to create SPASE descriptions from external sources of information;
- Harvester to extract information from SPASE descriptions or from registries;
- Wrapper to convert or embed SPASE metadata into other descriptions or forms such as OAI;
- Correlator to divide an XML document into individual resource descriptions to enable a well-organized file system; and
- Refcheck to determine the validity of the references in resource descriptions such as Resource IDs and URL's.

Additional tools or aids are planned for the future such as the Java-to-XML Binding Mechanism (JAXB) and a SPASE Guidelines Document for assistance in writing the SPASE descriptions.

THE FUTURE

Support and services are intended to extend into the future with continued revision of the data model to new versions, improvement of the editors and associated stylesheets, more complete registries, etc. As mentioned above there is also a plan to create a guidelines document for the SPASE description creator as well as tutorial and reference documents. Services will also be extended wherever resources allow. The Resource Registry is already in existence and will be improved in scope and availability. A versatile search engine across the heliophysics domain is an obvious use of the compilation of information acquired. A Representational State Transfer (REST) approach might also be used. Data visualization aids may be a possible added service as well.

In conclusion, we invite all in the space and solar physics domain to participate in SPASE. If you have resources of any kind whether they are data, event lists, or other items of interest that you want to share with the community, please contact any of the participating institutions. Those with a large volume of data or information to share are encouraged to set up their own virtual observatory attached to the SPASE domain. The rest of the SPASE community will be happy to help you and to build upon their experiences.

Domain specific data models and ontologies are making it possible for the seamless exchange of data across groups, agencies and international boundaries. With sufficient support (parsers, services, etc) adoption can be easy. With the right documentation mastering SPASE is possible to all. SPASE is working towards this goal. For more information go to http://www.spase-group.org.

REFERENCES

[1] -

[2] -

The authors would like to thank the many participants around the world who have contributed to the SPASE effort. Thanks also to the continued funding support from NASA which helps to enable a unified approach within the community.