The status of VxOs and their role in providing access to spacecraft data and promoting science

Aaron Roberts
NASA GSFC
12 May 2011
Decadal Review Working Group on Theory, Modeling, and Data Exploitation

The HP Data Policy Context

- The *HP Virtual Observatories* are one aspect of the HP Data Environment, as defined in the HP Science Data Management Policy (see http://hpde.gsfc.nasa.gov). The Data Policy has been in place since June, 2007, and has successfully guided the new missions, Data Centers, and others in matters of data preparation, storage/archiving, discovery, and access.

- Goals of the HPDE:

  - Produce and serve high-quality, well-documented data

  - Provide open access to scientifically useful data products

  - Keep data flowing without interruption when missions end

  - Keep data safe for the long term

What is a VxO?

- It is NOT:
  - A Data Center or Data Node
  - A Data Repository
  - A direct deliverer of data
  - *A website*
- It IS:
  - A service that ensures that all resources from subfield “x” are known, discoverable, and easily accessible. It looks to the user like a uniform data provider, but it is virtual.

What is the goal of VxOs?

- Enable science through efficient access to a wide range of resources and services.
  - Easy data finding and simple data access without having to do reformatting and writing input routines is already a big saving.
  - Such basic access is the foundation of other services such as data mining, correlation across data sets, use of event/feature lists for searches and direct studies, etc.

How does a VxO Work?

- VxOs use data product descriptions in standard Data Model terms to link users to repositories through “middleware.”
- Uniformity in protocols for accessing data from the repositories will greatly aid the process, although some non-uniformity will be inevitable and is workable.
- A VxO is only as good as its metadata.

VxO requirements

- Repositories (Mission, Resident, and Final Archives are working well).
- Uniform naming with a Data Model (SPASE is now a stable standard).
- Uniform format highly desirable (becoming true with FITS, CDF, and NetCDF), but not essential.

- Long-term access not part of VxO role, but data and metadata standards make archiving easier.

Heliophysics VxOs

- HP VxOs have “x” = Solar, Heliospheric, Magnetospheric, ITM, Radiation Belt, Wave, Energetic Particle; also a model-data comparison VO is called the Virtual Model Repository (VMR).
- There already is considerable reuse of tools and infrastructure in these.
- Further consolidation may be useful, but the diversity was to cover all areas for initial product descriptions and to possibly serve unique needs.

Heliophysics VxOs

- Each VxO has a website that delivers data or links to data, in nearly all cases with a search facility.
  - ViRBO has relatively fewer data products, ~30, and has focused on data upgrades of about 10 datasets that were not previously available to the community, rather than on a search engine.
- API access exists for most VxOs, with the same protocol (“SPASE-QL”) for half of them. All but VMR and ViRBO currently work on a “search then download resulting files” basis, and all have some browse capabililties.
- VITMO does a unique type of search, and includes a tool for finding conjunctions between spacecraft and ground-based observations.
- VHO and VMO/G allow the user to restrict the search based on statistical parameter values for selected variables in the files (e.g, parameter ranges).
- VEPO serves its products through VHO.
- VMO has two parts due to the number of datasets to be described (UCLA and UMBC/GSFC); these work together to avoid duplication.
- VHO, VMO/G, VEPO, and VWO share most of their core software for the middleware.
- VSO is the oldest VxO, the most mature, and the most integrated with the missions and archives (SDAC).

Heliophysics VxOs

- The Virtual Space Physics Observatory (VSPO) functions as an “active inventory” of all HP datasets
  - search and basic access capabilities as well as orbit and other services.
  - quick access to browse plots for many missions and instruments.
  - now part of SPDF, and represents one of many ways that SPDF is helping with the VO approach. [Others: providing RESTful, OPeNDAP, SPASE-QL, and SOAP access to its services; developing SPASE descriptions of resources; and working with Autoplot developers to improve CDF visualization and CDAWeb utility.]
- VxOs were started as largely independent SR&T projects, and have since been brought together within the Heliophysics Data and Model Consortium (see http://hpde.gsfc.nasa.gov).

- All VxOs are functioning and delivering data or pointers to it. The amount of data “delivered” and the number of papers that are based on it are still very much smaller than “conventional” routes in all cases except for VSO.

CDAWeb data directly into IDL: A prototype for direct data access from applications

- Download the file “spdfcdas.sav” [contains all needed CDF routines]
- In IDL:
  - restore, “spdfcdas.sav”
  - uly1sec95_6 = spdfgetdata('UY_1SEC_VHM', ['B_RTN', 'B_MAG'], ['1995-06-29T00:00:00.000Z', '1995-06-31T00:00:00.000Z'])
  - structid = ‘uly1sec95_6’ [the name of the ‘structure’ with everything in it]
  - assign_variables [Pulls the variables out of the structure and gives them names according to the CDF metadata]

  - [Can also invoke the following for a gui dataset/variable/time range chooser:]
  - spdfcdawebchooser [allows direct reading or command generation]
  - [IDL reports back the names of the variables that have been read in.]

  - qq = plotmaster(uly1sec95_6,/auto) [will plot all the data as in CDAWeb]
  - [A general routine exists to put variables on a uniform time basis by averaging or interpolating as needed.]
- This approach can be generalized using SPASE descriptions that give IDs and Parameter Keys for a broad range of datasets, along with and a standard protocol for access. (Currently being developed based on OPeNDAP.) VSO also does this now, using its data model.

Answers to the Working Group’s Questions

Available datasets

- Any dataset with a SPASE description is “available,” although it nearly always actually comes from a data center or a mission repository. A relevant metric is the fraction of data products that have SPASE descriptions, based on an inventory we are compiling. Overall, this is roughly 85% of all HP resources for basic descriptions, and ~50% for detailed (parameter level) descriptions.
- The basic descriptions of and associated access to data products are now provided for all active and most past NASA HP missions through VSPO. VSO provides (nonSPASE) access to ~85% of all solar resources (by number), and nearly 100% by data volume (i.e., SDO, STEREO, SOHO, Hinode, and major ground-based facilities). SPASE-based searches for and access to solar resources (via pointers to URLs or to VSO) are provided via VSPO for all the major NASA missions, and this includes access to browse products such as LASCO movies at NRL.
- Direct file access is increasingly provided by VxOs as the detailed SPASE descriptions are completed.

Visualization software

- Autoplot: an open-source 2-D plotting application, reads a wide variety of formats of files, allowing users to graphically explore data. It has a server-side version that can be used by a VxO for browse plots, as ViRBO does. It started with ViRBO, but many of the VxOs and some of the RBSP teams have adopted Autoplot.

- ViSBARD: an open-source 3-D visualization application that allows the user to view measured vectors and scalars along the spacecraft orbit. It includes the ability to generate Tsyganenko field lines, model bowshock and magnetopause boundaries, and now allows some inclusion of simulation output from CCMC. User-base is not known.

- TIPSOD: A 3-D orbit viewer from SPDF is increasingly used by VxOs as a Java webstart tool to allow 3-D orbit viewing. Considerable use in the community.

Number of users

- It is difficult to determine who are “real users” as opposed to those just coming to window shop, but the number of unique IPs recorded by the VxOs varies from around 300 to 2000 per month; this number increases when the integration time is increased. VHO and VMO/G now track the number of actual queries executed, which comes to about 100/month each. VSPO is visited by ~700 unique IPs per month, and has ~10k downloads of search pages, but more analysis is needed to see what this means in terms of real searches for datasets. As a point of reference, SPDF has about 38k “executions of software” (e.g., make a plot or a dataset) per month.

Volume of downloaded data

- Since VxOs mostly deliver metadata, not data, this is difficult to determine.

- The VxOs that serve some files from their machines provide users with ~20-60 GB/month. (The VSO-SDAC-SDO connection is an exception with a few TB/month. SPDF data volume is ~1.4 TB/month, including 600K files delivered via ftp.)

Number of papers published using data obtained from the VxOs

- It is somewhat difficult to know who has used a VxO to obtain the data they use. A search for any use of the full names or abbreviations of VxOs in JGR in GRL papers for 2010 yielded 6 hits for VITMO and 5 for VIRBO. One VHO and one VITMO hit were found in 2009. (VSO hits were not expected.) SDAC reports ~100 papers/year using SDAC resources, but VSO is not singled out. (We will do a specific search in solar journals.) SPDF data services and datasets are mentioned in ~180 papers/year in JGR and GRL (~20% of total papers).

Mechanisms that have been put in place to facilitate feedback between the leadership of the VxO and the scientific user community

- All the VxOs present at large meetings such as the AGU. VxOs work directly with data providers to write or improve SPASE descriptions of resources. For the new missions, VSO has worked with all the recent solar missions to provide access. VMO is working with MMS and ViRBO is working with RBSP to (when the time comes) provide SPASE descriptions of data products and to assure easy general access to them.

Areas of the HP Data Environment That Need Continued Support

- Easy browser and direct machine access to reliable, well documented, well preserved data sets.
- A complete and useful inventory/registry of data products and services.
- Format and protocol standards to make everything else easier.
- As mentioned above, these will form the foundation for services such as “dataset runs on request,” data mining, use of event/feature lists, parameter value constrained searches; etc.

VxO Contributions

- The VxOs have made substantial contributions in all the above areas through:
  - defining the SPASE Data Model,
  - creating SPASE descriptions along with registries and tools for using them,
  - providing file-level data access, and working to develop protocol standards,
  - contributing generic tools such as Autoplot.
- VxOs have contributed to the quality of data documentation and in some cases to the “upgrading” of data products for general use.

- More complete inventories of SPASE descriptions will increase user confidence in being able to find data, and better services that allow, for example, the downloading of data in desired formats should improve usage.
- We must determine if VxOs are just young, or if novel approaches to implementing or advertising the tools are needed to improve actual and perceived utility.

Longer-term outlook

- Long-term, VxOs will have a role in assuring the completeness of the HP data inventory and in providing complete SPASE descriptions of all resources, in particular working with new missions to provide high-quality, complete metadata.
- They will be essential in defining and implementing protocols that will allow easy access to data from applications such as IDL or Python, as now done widely in the Earth Sciences and Astronomy, and as we are starting to do.
- It is less clear what the role of specific web portals will be, but we should be able to determine this over the next year or two. It may be that fewer, more capable portals are needed, or some things that are now portals (e.g., the VHO/VMO data selection by parameter range) could become services used by other software.
- It remains to be seen at what level browser VxO access will supplant the already quite successful access to Mission, Resident, and Final Archives. My own guess is that web browser access in general will (very?) gradually fade in favor of more direct access by applications, services, and tools; the VxO infrastructure will facilitate this.

Extra slides

- Presentation to the HP Data and Computing Working Group, 23 Feb 2011

The Data Policy states various functions for the HPDE

- Produce and serve high-quality, well-documented data

- Provide open access to scientifically useful data products
  - Allow easy discovery of all available products and their location
  - Provide easily useable, well-documented products
  - Provide uniformity of access to data

- Keep data flowing without interruption when missions end
  - Provide funds to continue post-mission serving of data
  - Move data to Active Final Archives for long-term serving

- Keep data safe for the long term
  - Assure data are safe at all stages
  - Provide long-term archives for safe-keeping

The HP Data Policy is working

- New missions are following PDMP guidelines and will deliver data as expected; VxOs and Final Archives are involved in the process.
- Current missions are improving their data, documentation, and services; most are in good shape.
- Senior Reviews and Mission Archive Plans continue to help.
- Data are moving into Active Final Archives, and are being served and kept safe.
- An Inventory and Registry of all HP data is being completed and has an active interface (VSPO) that will deliver or point directly to data.
- Legacy datasets are being improved, archived, and served.
- Plans are moving forward for uniform access to HP data.
  - HDMC/VxOs

Inventory/Registry: SPASE is stable and working

- Most data products from nearly 100 space-based and many more ground based observatories are registered using SPASE (includes 30 solar observatories, space- and ground-based)

- Nearly all available data from all NASA HP active missions is directly accessible
  - Easily discovered by time range, cadence, general region, measurement type, name, relevant text in description, person name, ... or any combination of the above.
  - Parameter range, magnetospheric state, spacecraft/ground coincidence, and/or event lists available for searching for some data products, depending on VxO.

- Non-NASA data largely accounted for (some availability and access problems)

- Lag in SPASE descriptions at the detailed (parameter/variable) level
  - Affects universality of access and limits some types of search
  - Being addressed

Problem of Uniform Access (asking for all data in the same way)

- Advantages of self-documenting, standard formats
  - Variable names, units, etc., are encoded in a uniform way
  - Time is in a fixed format and is thus instantly readable
  - Descriptive metadata is tied to the relevant variables
  - Internet access from, e.g., IDL or MatLab can be easily automated
  - *CDF-A (CDF + time, structuring, and metadata conventions) being developed to have a truly archival CDF*
  - *FITS and NetCDF (probably TIMED conventions) should complete our set*

- SPASE-based access (e.g., access by: SPASE ID; time range; variable 'keys')
  - Metadata required, but can be difficult to get (progress being made)
  - SPASE-QL and/or general Data Access Protocols use the metadata
  - Currently implemented by some VxOs and CDAWeb.
  - VxOs—possibly ultimately not so much portals as formulators and implementers of standards (VAO/IVOA path) for the protocols; general tools build on that.

- ASCII problem
  - Lack of standards means more metadata required
  - Schemes for generating and using such metadata are being generated
  - Copies in standard formats can and do solve the problem

Most datasets are now safe for the long-term and actively served

- Science-quality, high-resolution data are at SPDF (CDAWeb or ftp), in most cases for most or all instruments:
  - ACE, Wind, Polar, IBEX, Voyager, Pioneer, Helios, THEMIS, STEREO (in situ), Ulysses, SOHO (particles), Geotail, IMP-8, DE, Many Explorers, ISIS, TWINS, Cluster (prime parameters), some others being negotiated. Also OMNI.
  - Other countries also preserve data (notably, the Cluster Active Archive, but also many others such as Akebono and Geotail at DARTS).
  - RAs keep IMAGE/RPI, FAST, many Polar datasets, and other space physics data flowing.
  - Long-term backups via NSSDC

- There are many active solar missions (Hinode, SOHO, SDO, RHESSI, STEREO imaging); data are well served and probably quite safe, e.g., with copies served from SDAC, but not as clear a plan in some cases. RAs exist for a number of older missions (TRACE, Yohkoh, SOHO MDI), and other countries also preserve data (e.g., Hinode at DARTS; SOHO and RHESSI in Europe).

- Probably safe, but no NASA plan: IMAGE ENAs, SAMPEX, non-NASA (DMSP, NOAA, etc.)

AGU statement as an indication of community agreement on data archiving

- "Documenting trends and long-term changes is essential for understanding many natural phenomena. Because the state of natural systems is never repeated, data losses, or missed data collection opportunities can never be corrected. Consequently, *the value of Earth and space science data grows with time*, placing a premium on *long-term data curation*. Because datasets are often later used for purposes other than those for which they were collected, *accurate, complete, and, when possible, standardized metadata are as important as the data themselves."*
  - AGU Position Statement on **The Importance of Long-term Preservation and Accessibility of Geophysical Data**
  - http://www.agu.org/sci_pol/positions/geodata.shtml

Datasets being restored/improved/upgraded

- ISEE-1, -2, FAST, WIND/SWE, SUSIM irradiance, Mees Vector Magnetograms, DE-1 plasma waves, SMM Gamma-rays, a few others

- We are reaching the end of the list of useful cases
  - New proposals tend to be for more subtle improvements rather than basic restoration.
  - Remaining known datasets (e.g., at NSSDC) currently in nonstandard form are typically older, shorter, and “less interesting.”
  - There may be some things we just cannot afford although they would be useful, but not many.

Future challenges/vision

- Metadata production and use
  - Definitive inventory/registry: referential (DOI?) and discovery uses
  - Uniform data access for all products
  - Seamless flow from mission archives through to final archives

- Format standards (e.g., CDF-A; also NetCDF standard?)
  - Adoption of standards in calls for mission proposals (the time has come)

- Large data volumes
  - How to use the data: Pattern recognition; data mining
  - How to keep the data available and safe post-mission

- Model-data comparisons and insights
  - Seamless integration of model output with data streams
  - Data assimilation; true space weather capabilities
  - Data volume questions, as above

Future challenges/vision (Decadal Survey White Paper)

- Heliophysics science requires efficient, long-term access to well-maintained repositories of carefully prepared, documented, and preserved data to "develop an integrated research strategy that will present means to address [high-priority scientific] targets," as required in the charter for this Decadal Survey. Because of this, we strongly urge the decadal committee to:
  - reaffirm support for the "Solar and Space Physics Information System" recommended in the last decadal report, and that has been moving forward due to the efforts of many countries and agencies;
  - assert the importance to the accomplishment of science goals of adequate and sustained funding for agency efforts to maintain and further improve long-term archive and distribution mechanisms; and
  - strongly support the need for general standards for archiving and distribution of data as exemplified by those contained in the NASA Heliophysics Science Data Management Policy, and starting with the endorsement of an open data policy by all relevant agencies.

Future challenges/vision

AGU statement as an indication of community agreement with the White Paper assertions

- "AGU policy is grounded in the principle of *full and open sharing* of [space and Earth science] data and associated metadata for research and education. Adherence to this policy will foster scientific advances, yield economic benefits, improve decision-making, enhance public safety and wellbeing, contribute to national and global security, and lead to a more informed public."
- *"The cost of collecting, processing, validating, and submitting data to a recognized archive should be an integral part of research and operational programs. Such archives should be adequately supported with long-term funding.* Organizations and individuals charged with coping with the explosive growth of Earth and space digital data sets should develop and offer *tools to permit fast discovery and efficient extraction of online data, manually and automatically*, thereby increasing their user base. The scientific community should recognize the professional value of such activities by *endorsing the concept of publication of data*, to be credited and cited like the products of any other scientific activity, and *encouraging peer-review* of such publications."
  - AGU Position Statement on **The Importance of Long-term Preservation and Accessibility of Geophysical Data**
  - http://www.agu.org/sci_pol/positions/geodata.shtml