Status and Vision for the Heliophysics Data Environment (HPDE)

Aaron Roberts
NASA GSFC
23 February 2011
HPDCWG

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

Extra slides

A Long-Time User's Expectations (McPherron's AGU talk, 2009)

1. A long list of URLs for a variety of data systems each with different interfaces and rules will not be needed
2. The VMO system will be simple to use and can be quickly learned by trial and error without recourse to complex documentation
3. The system will NOT have multiple pages with innumerable fields
4. It will be easy to determine whether data of a particular type exists for the interval of interest
5. The system will deliver data in a specified format that is convenient to work with
6. The system will not assume a specific method of research
7. VMO will not be an analysis or publication graphics system
8. Data transfer will be simple regardless of quantity

Citing Data in Regular AGU Journal Papers

"Data sets cited in AGU publications must meet the same type of standards for public access and long-term availability as are applied to citations to the scientific literature. Thus data cited in AGU publications must be permanently archived in a data center or centers that meet the following conditions:

a) are open to scientists throughout the world.
b) are committed to archiving data sets indefinitely.
c) provide services at reasonable costs.

The World and National data centers meet these criteria. Other data centers, though chartered for specific lengths of time, may also be acceptable as an archive for this material if there is a commitment to migrating data to a permanent archive when the center ceases operation. Citing data sets available through these alternative centers is subject to approval by AGU."

- **Policy on Referencing Data in and Archiving Data for AGU Publications**
- http://www.agu.org/pubs/authors/policies/data_policy.shtml

CDAWeb data directly to IDL

- 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.]