An Overview of the Space Physics Data Facility (SPDF) in the Context of “Big Data”

Bob McGuire, SPDF Project Scientist
Heliophysics Science Division (Code 670)
NASA Goddard Space Flight Center

Presented to the Big Data Task Force, June 29, 2016
Topics

- As an active Final Archive, what is SPDF?
  - Scope, Responsibilities and Major Elements
- Current Data
- Future Plans and BDTF Questions

REFERENCE URL:  http://spdf.gsfc.nasa.gov
One of two (active) Final Archives in Heliophysics
- Ensure the long-term preservation and ongoing (online) access to NASA heliophysics science data
  - Serve and preserve data with metadata / software
  - Understand past / present / future mission data status
  - NSSDC is continuing limited recovery of older but useful legacy data from media
- Data served via FTP/HTTP, via user web i/f, via webservices
- SPDF focus is non-solar missions and data

Heliophysics Data Environment (HpDE) critical infrastructure
- Heliophysics-wide dataset inventory (VSPO->HDP)
- APIs (e.g. webservices) into SPDF system capabilities and data

Center of Excellence for science-enabling data standards and for science-enabling data services
SPDF Services

• **Emphasis on multi-instrument, multi-mission science**
  
  (1) Specific mission/instrument data in context of other missions/data
  
  (2) Specific mission/instrument data as enriching context for other data
  
  (3) Ancillary services & software (orbits, data standards, special products)

• **Specific services include**
  - Heliophysics Data Portal (heliophysics-wide “find data” and services”)
  - CDAWeb (multi-mission data browsing and correlative analysis)
    - Now: Wind, ACE, THEMIS, STEREO, VAP, MMS, many ground stations
    - Coming: ICON, GOLD, SPP (in-situ)
  - SSCWeb & 4D Orbit Viewer (satellite orbits and ground tracks)
  - OMNIweb Plus (solar wind parameters mapped to nose of bow shock)
  - CDF (Common Data Format) plus structure/metadata guidelines
    - Self-describing data format, scalar and multidimensional data, platform- and discipline-independent, common model with netCDF
    - Implementation guidelines for organization, naming, metadata
    - Use of “master” CDFs to update/add metadata and capabilities
CDAWeb Data Explorer

- Select start and stop times from which to GET or PLOT data:
  - Use pre-defined start/stop times
    - September 2005 Events: 09/01-09/30: 09/01 - 09/31
    - Use custom start/stop times
      - Start: 2012/10/02 00:00:00 (YYYY/MM/DD HH:MM:SS mm:ss)
      - Stop: 2012/10/03 00:00:00 (YYYY/MM/DD HH:MM:SS mm:ss)
  - Select an activity:
    - Plot Data: select one or more variables from list below and press submit.
      - Also create PS and PDF outputs (all plot types except images and plasmagrams).
      - CDAWeb supports the following data formats:
        - HDF5 (HDF5, NS2000, and NSPOD), NetCDF (NCDF), ASCII, and CDF
      - Note: CDF files are created using Version 3.4 CDFS in IDL or MATLAB.
    - Plotting Options:
      - Use source noise filtering to remove values outside 3 standard deviations from mean of all values in the plotted time interval.
      - Double-click the Y-axis value for time-series and spectrogram plots.
      - Combine all time-series and spectrogram plots, for all requested datasets, into one plot file.

Missions and Instrument Types

- ACE
- ARTEMIS
- BARREL
- CROPS
- CRES
- Cluster
- Cubesats
- EBSM
- Equinox-5
- FAST
- GOES
- GPS
- Geomag
- Geotail
- IMAGE
- IMP-A (All)
- IMS
- Intahs
- LANL
- MESSENGER
- MMS
- NOA
- OMNI
- NOAA
- POES/MetOp
- Pioneer
- Poly
- RODOSAT-1/FYB/SAT-1/YPE
- SAMPEx
- SNOG
- STS
- STEREO
- THEMIS
- TIMED
- TVVS
- Ulysses
- Van Allen Probes (RISP)
- Voyager
- Wind
- Ground-Based Investigations

Sample Plots of GOES and THEMIS data
- Also ASCII and CDF downloads, with super/subsets
- All these services are also available through RESTful webservices API

G10_12_MAG
GOES-10 High Resolution Magnetometer data (at 512 samples/sec at 60 sec) - Howard J. Singer (NOAA Space Weather Prediction Center)
Available dates: 2007/08/01 00:00:00 - 2007/08/31 23:59:59

G11_12_MAG
GOES-11 High Resolution Magnetometer data (at 512 samples/sec at 60 sec) - Howard J. Singer (NOAA Space Weather Prediction Center)
Available dates: 2008/01/01 00:00:00 - 2008/12/31 23:59:59

Variables
TIMED GUVI Data on CDAWeb


SSCWeb & 4-D Orbit Viewer

RBSP and THEMIS
Northern Hemisphere
Animated Ground Tracks
in December 2012

MMS, Van Allen Probes,
Cluster and Geotial orbits
December 12, 2015
FY16 Mission and Data Highlights

- **Ingest/serve data from operating heliophysics missions**
  - Now serving publicly released MMS L2 CDF data (50x4) and QL plots
  - Continuing to add Van Allen Probes L3 datasets (plus ongoing data flows)
  - Continuing archival data captures from AIM and TIMED (netCDFs)
  - New datasets from active missions, including e.g. Geotail CPI and EPIC

- **Other ongoing acquisition**
  - Past missions: E.g. Polar VIS, IMAGE, FAST, Tether
  - Other missions (NASA heliophysics relevant):
    - GOES/POES particles-magnetic field (w NOAA)
    - DMSP SSUSI (w APL), SSJ (w NOAA), SSM (w U. Colorado)
    - Orbit information for SSCWeb database

- **Ingesting and serving older data as available from NSSDC**
  - Unpacking and organizing, limited improved metadata

- **Regular interactions and planning with upcoming missions**
  - Presented at several GOLD team meetings; telecons with ICON
  - Regular participation in the SPP and SPP SOC meetings

- **Software and System to support ingests, new data, new capabilities**
  - Note CDF s/w and CDAWlib IDL library are open source
~33% of 2015 papers in AGU’s JGR Space Physics acknowledged SPDF services and/or data (acknowledgments were ~ 20-25% in 2009-2014)
Dominant 2016 data ingests are MMS and Van Allen Probes
- GOLD data rate starting in 2018 will be comparable to MMS
- 1st Quarter 2016 data downloads ~20 Tbytes/~1.4M files per month
## Missions Supported in SPDF

<table>
<thead>
<tr>
<th>ACE</th>
<th>GOES 6-15</th>
<th>OGO 1-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>GPS</td>
<td>Pioneer 6-11</td>
</tr>
<tr>
<td>Alouette 1/2</td>
<td>~120 Ground-Based Obs</td>
<td>Polar</td>
</tr>
<tr>
<td>AMPTE</td>
<td>Hawkeye</td>
<td>ROCSAT</td>
</tr>
<tr>
<td>Apollo</td>
<td>Helios 1/2</td>
<td>S-Cubed</td>
</tr>
<tr>
<td>Arcad</td>
<td>IBEX</td>
<td>SAMPEX</td>
</tr>
<tr>
<td>Ariel</td>
<td>IMAGE</td>
<td>San Marco</td>
</tr>
<tr>
<td>ATS 1-6</td>
<td>IMP 1-8</td>
<td>Skylab</td>
</tr>
<tr>
<td>BARREL Balloons</td>
<td>INJUN</td>
<td>SNOE</td>
</tr>
<tr>
<td>Cluster</td>
<td>Interball</td>
<td>STEREO</td>
</tr>
<tr>
<td>CNOFS</td>
<td>ISEE 1/2/3/ICE</td>
<td>THEMIS/ARTEMIS</td>
</tr>
<tr>
<td>CRRES</td>
<td>ISIS1/2</td>
<td>TIMED</td>
</tr>
<tr>
<td>DE 1/2</td>
<td>LANL 1989-2002</td>
<td>TWINS</td>
</tr>
<tr>
<td>DMSP</td>
<td>MAGSAT</td>
<td>Ulysses</td>
</tr>
<tr>
<td>Explorer 4-35</td>
<td>Mariner 4-9</td>
<td>Van Allen Probes</td>
</tr>
<tr>
<td>FAST</td>
<td>Messenger</td>
<td>(RBSP)</td>
</tr>
<tr>
<td>Genesis</td>
<td>MMS</td>
<td>Voyager 1/2</td>
</tr>
<tr>
<td>Geotail</td>
<td>NOAA 5-19</td>
<td>Wind</td>
</tr>
</tbody>
</table>
The Complexity / Heterogeneity Dimension

• Missions and data cover a wide scope of science problems and measurement techniques
  – In-situ and remote sensing
  – Interplanetary to interstellar space,
  – Earth’s magnetosphere

• >1400 individual datasets and ~30M files
  – With >25,000 individual parameters (just for the data now in CDF)

• Mission/community acceptance of data standards is key
  – Support standards with s/w and guidelines to enable serving data
  – Work with the missions in using these standards
  – Serving the data as an inherent value to the missions
  – Immediate capture of mission data best ensures long-term preservation and quality control by the community
SPDF Technical Architecture

External Science Community

SPDF Public-Facing Web Server (with Multiple Virtual Hosts)

Private Switch

Development and Management Server

Internal Staff

Storage Servers (Mirrored RAID, each w 250 TB usable)

Backup Tape System
A Status and Planning Summary

• **SPDF is a team effort**
  – Local SPDF scientists, software professionals and support staff
  – And … our mission data provider and research user communities

• **SPDF will be able to manage currently expected data volumes**
  – Direct work with missions and early capture of data are both very important
  – Safe copies of data are stored remotely from Goddard
  – SPDF has developed format converters e.g. CDF <-> netCDF

• **SPDF supports the wider Heliophysics Data Environment**
  – Standards + services + data available by webservices
  – Supporting infrastructure for analysis s/w such as Autoplot and SPEDAS

• **SPDF will consider cloud for next major storage increment**
  – e.g. single string our RAID storage and maintain second copy in a cloud
A Few Challenges and Other Notes

• **Technology continues to move rapidly**
  – H/W capabilities but also changing S/W development environment
    • What’s the future of IDL? What’s the future of Java?
    • Are there better stable options for long-term future development?

• **How should cloud technology support a long-term archive**
  – Need local copies for staging and to ensure preservation?
  – File organization (S3 vs. current disk directory organization)
  – Fedramp compliance can easily be a cost concern

• **Archiving**
  – Long-term approach to older data
    • All public SPDF data are kept online and accessible, even where difficult to use
  – SPDF current scope is observational data and not model outputs
  – Better to do “something”
(1) What are the processes for planning for future (5-10 years) capabilities of your service? How and from whom do you gather input for this planning process and where does input typically come from? What new features have highest priority?

• Starting point is regular/ongoing assessment of upcoming (approved) missions for expected data volume and unique features or requirements to ingest/preserve/serve that data
  – Very hard to go out beyond 5 years
• Regular telecons with various project, data and data services groups
• Staff attends/presents at science meetings regularly and brings back issues
• Ongoing process of technology assessment
• Periodic Senior Reviews or other program-level reviews

• Mission data and service requirements are highest priority
  – E.g. better support netCDF for the upcoming ICON & GOLD missions
  – E.g. expected GOLD data rate starting 2018 ~ same level as current MMS
(2) What feature(s) of your service would you like to stop performing? How do you gather input for making such decisions and where does input typically come from? What is preventing you from stopping?

- All present services appear to be useful and are being used, based on usage statistics and individual user feedback
- Stopping any major capability would be discussed with HQs program management and advertised to the research community
(3) **What steps you are taking to make your data interoperable with allied data sets from other data sites in and out of NASA? How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?**

- Highest degree of interoperability is obtained by holding data to standards being used by other NASA heliophysics missions, i.e. holding/serving data in CDF and following the ISTP/SPDF metadata.
- PDS is now accepting CDF (with a few specific extra constraints).
- First priority for ingest and preservation always goes to NASA Heliophysics mission and other NASA heliophysics data.
- Other relevant data is usually brought to SPDF’s attention by potential providers, because they like our services and want to be included.
- Other NASA SMD and other non-NASA data is assessed for NASA science relevance/importance, special issues of urgency (e.g. important data at risk of loss or not easily available to public) and impact to SPDF resources, in consultation with our NASA HQs program manager.