Importance of Heliophysics standards and metadata guidelines for effective data analysis

Robert M. Candey[1], Bernard T. Harris[2], Tamara J. Kovalick[3], Michael H. Liu[5], Robert E. McGuire[1], D. Aaron Roberts[1]

[1]Code 670/NASA Goddard Space Flight Center, [2]Code 580/NASA Goddard Space Flight Center, [3]ADNET/NASA Goddard Space Flight Center

Poster: IN11D-0651

AGU 100 FALL MEETING Washington, D.C. | 10-14 Dec 2018

Effective data analysis and browsing using generic easy-to-use software and web services are enabled by:

- Standardized self-describing data formats
- Standardized metadata for datasets and parameters
- Standardized time conventions
- Standardized dataset and filenaming conventions

Alternative is laborious and time-intensive custom code in every programming language for every dataset

Why metadata conventions

- Standardized self-describing data formats, metadata for datasets and parameters, time conventions, and dataset and filenaming conventions enable effective data analysis and browsing using generic easy-to-use software and web services
- Restricting metadata representations limits the number of equivalent possibilities with which software must deal, and thus fosters interoperability
- Conventions standardize ways to name things, represent relationships, and locate data in space and time
- Enables developing applications with powerful extraction, regridding, analysis, visualization, and processing capabilities
- Abstracts general data models to represent data semantics
- Embodies provider's experience and captures the meaning in data and make data semantics accessible to humans as well as programs
- Higher-level abstractions such as coordinate system and standard names for physical quantities enable comparing different data, and distinguishing between variables
- Standard data and metadata in modern formats enables migration to follow-on standards

Some numerical conversion utilities for reading old formats

- conv_vax_unix_v4.pro in IDL <https://spdf.gsfc.nasa.gov/pub/software/format_conversion/conv_vax_unix_v4.pro>
- FltPnt Ruby routine for converting most everything <http://float-formats.rubyforge.org/classes/FltPnt.html>
- Univac format <http://www.fourmilab.ch/documents/univac/minuszero.html>
- Old formats <http://nssdc.gsfc.nasa.gov/nssdc/formats/>

Scientific data file formats and standard metadata in NASA Space Science

- **FITS** used in astronomy and solar physics [FITS and WCS metadata]
- **HDF** in Earth sciences [HDF-EOS hdfeos.org metadata]
- **netCDF** in atmosphere [Climate and Forecast cfconventions.org] and ITM [ISTP/SPDF metadata]
- **CDF** in the rest of Heliophysics [ISTP/SPDF Guidelines metadata]
- **PDS** (and JPEG) in planetary [PDS metadata]; recently added **CDF-A** as standard format (CDF with ISTP/SPDF Guidelines and two SPASE attributes, but no compression or sparse variables)
- CDF/netCDF compatibility: netCDF4 Classic model with no groups or user-defined variable types, time should be unlimited dimension
- SPDF converters between CDF, CDFML, netCDF, HDF-4, FITS, and to PDS-3

Heliophysics standards and conventions developed through hard experience

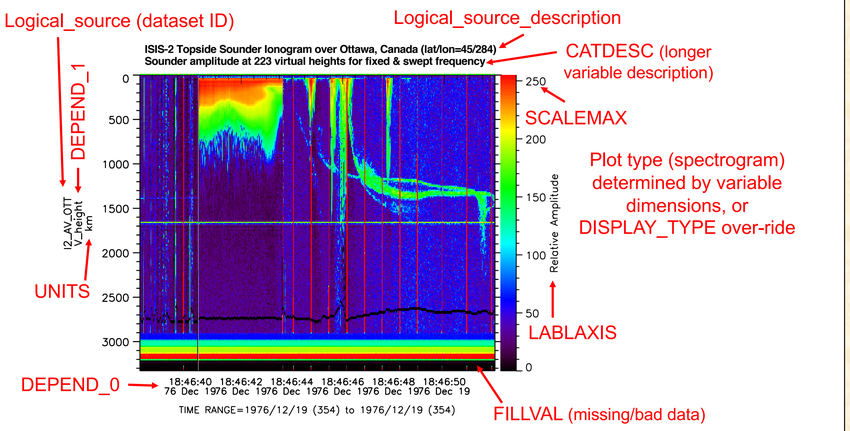
- **SPASE** <http://www.spase-group.org> dataset descriptions for easy searching
  - **Heliophysics Data Portal** <https://heliophysicsdata.gsfc.nasa.gov>
- **ISTP/IACG/SPDF Guidelines** for global and variable attributes <https://spdf.gsfc.nasa.gov/sp_use_of_cdf.html>
  - **SKTeditor** metadata creation tool <https://spdf.gsfc.nasa.gov/skteditor>
- **Dataset naming and file naming** recommendations <http://www.tsds.org/Recommended_file_and_data_collection_naming_practices>
- **Filenaming templates** <http://tsds.org/uri_templates> $Y/data_$Y_j_idx.cdf
- **CDF** <https://cdf.gsfc.nasa.gov> scientific data format (including its new Python library <https://github.com/MAVENSDC/cdflib>)
  - Time variable types <https://cdf.gsfc.nasa.gov/html/leapseconds_requirements.htm>
- **netCDF** <https://www.unidata.ucar.edu/software/netcdf/>
- **FITS** <https://fits.gsfc.nasa.gov/>
- **Heliophysics Event List** (Catalog) format <http://spase-group.org/docs/conventions/HDMC-Event-List-Specification-v1.0.3.pdf>
- Some tools enabled by these standards:
  - **CDAWeb** <https://cdaweb.gsfc.nasa.gov> and **CDAWlib** IDL library
  - **Autoplot** <http://autoplot.org>
  - **SPEDAS** <http://spedas.org> IDL library

CDF standard time variable types

- **CDF_TIME_TT2000** nanoseconds from J2000 in Terrestrial Time in 8 byte integer handles leap seconds and is well-defined; UTC conversion requires up-to-date leap second table (last value stored in CDF header as a check)
- **EPOCH** milliseconds from 0AD in 8byte float; usually UTC but not leap seconds
- **EPOCH16** picoseconds from 0AD in two 8byte float; usually UTC but not leap seconds

Time variable types <https://cdf.gsfc.nasa.gov/html/leapseconds_requirements.html>

Self-describing datasets and ISTP Metadata provide logical/semantic structure for automated processing



SPDF Services enabled by these standards

- **Archive** for non-solar NASA Heliophysics science data and many other missions (spdf.gsfc.nasa.gov/pub/)
- **CDAWeb** browse, correlations and display, simple interface
- **SSCWeb** orbit/ground track data and conjunction queries, 4D viewer
- **OMNI Data/OMNIweb-Plus** (baseline solar wind data at Earth)
- **Heliophysics Data Portal (HDP)** SPASE-based inventory of public Heliophysics-relevant data
- **CDF** self-describing scientific data format
- **SKTeditor** for creating and testing **ISTP/SPDF Guidelines** metadata (CDF/netCDF)
- **Master** CDF/netCDF concept uses file with no data to add/over-ride metadata in datasets
- **Web services** for CDF/netCDF data in CDAWeb, SSC orbits, OMNIweb, HDP; use REST versions, many language examples <https://cdaweb.gsfc.nasa.gov/WebServices/REST/> (same for SSCweb)
- **HAPI** interface to CDAWeb holdings <https://cdaweb.gsfc.nasa.gov/hapi>

All SPDF Data and Services can be reached at spdf.gsfc.nasa.gov