ISTP Metadata Guidelines Overview

Robert Candey on behalf of SPDF Team

Space Physics Data Facility (SPDF)
https://spdf.gsfc.nasa.gov
Heliophysics Science Division (Code 670)
NASA Goddard Space Flight Center

IHDEA Meeting 2023 Oct. 13

Why Metadata Conventions

- Leverage standardized self-describing data formats, metadata for datasets and parameters, time conventions, and dataset and filenaming conventions to 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
- Embody data provider’s knowledge and capture the meaning in data and make data semantics accessible to humans as well as programs
- Provide higher-level abstractions such as coordinate systems, standard names for physical quantities for comparing different data and distinguishing variables

ISTP/SPDF Guidelines Structure and Metadata Concepts

- **ISTP/IACG Guidelines (mid 1990s) and subsequent extensions by SPDF define implementation standards for CDFs and NetCDFs**
  - Include general file naming conventions
  - Data is time-ordered and time-identified; times vary by record
  - Set of required and suggested metadata (details on next slide)
  - Variable attributes can point to other variables by name and carry arguments
    - Attributes thus carry information about relationships among variables
    - Variables can carry metadata (e.g., labels for dimensional variables)
  - Global attributes provide overall context of the dataset
  - Missions add their own metadata requirements

- **CDAWeb additional concepts: "Master" CDFs and "Virtual" Variables**
  - "Master" CDF is the use of a "skeleton" CDF (structure and metadata but no data) to insert supplemental or updated metadata for CDFs as a dataset
  - ”Virtual" variables are computed variables, using specialized CDF attributes to link defined variables and routines within CDAWeb/CDAWlib

ISTP/SPDF Metadata Elements

- **Variable attributes required for automated processing:**
  - Catdesc for longer variable description
  - Depend_0 points to time variables
  - Depend_1, 2, 3 point to variables that describe other dimensions
  - Fieldnam short variable name for plots
  - Fillval values indicating missing or bad data
  - Lablaxis/Labl_ptr for axis and column titles
  - Units/Unit_ptr
  - Validmin/max for valid data range
- **ISTP metadata independent of CDF and easily used in other self-describing science formats like CEFs, netCDFs and HDFs, and probably FITS and ASDF**
- **CDF 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 8-byte float; usually UTC but not leap seconds
  - EPOCH16 picoseconds from 0AD in two 8-byte float; usually UTC but not leap seconds

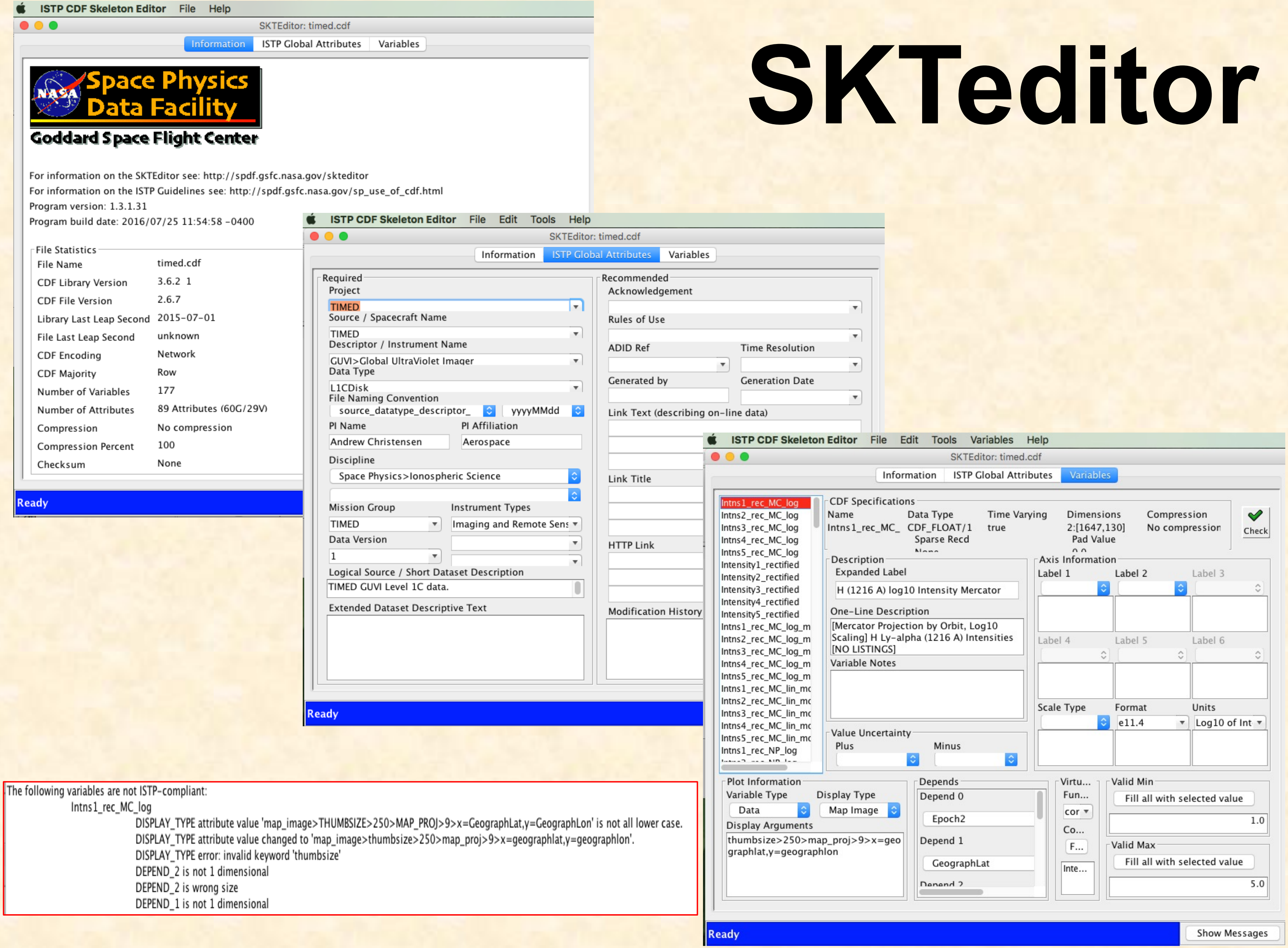
Development

- Converted Guidelines to Markdown format and added to *https://github.com/IHDE-Alliance/ISTP_metadata*
- Bobby is still editing to add changes identified in the past few years, and adding general dataset creation recommendations and lessons-learned
- **Added some global attributes and variable attributes** to ISTP standard, such as author list for DOIs, DOI, Variable_display_order, Variable_display_indent_level, Associated_parent_variable, Dataset_group, Mission_parent) and from Cluster/Solar Orbiter: Representation, Tensor_order, Coordinate_systems, Rotation_matrices, Unit_quarternion
- Consider requirements specific to model results
- The Earth science community uses the CF Conventions (originally Climate and Forecast) *https://cfconventions.org/*
- Future governance might be overseen by an international committee or fold into the SPASE effort
- Better document Guidelines on Github with mission-specific metadata as well, but want to keep flexible for interactions with missions and enabling framework for CDAWeb services

Tool to Create/Edit a CDF/NetCDF File Compliant to ISTP/SPDF Standard

- SKTeditor is a Java, web-start application, soon to be in JavaScript
  - Guide designers to good choices consistent with ISTP/SPDF guidelines
  - Create new CDF/NetCDF or check/correct then modify existing skeleton file
- Guided by the interface flow, add or edit
  - Scalar and higher-dimensional variables, multiple time variables
    - Times as cdf_epoch or preferably cdf_time_tt2000
  - Variable attributes (descriptions, labels, units, display_type)
  - Global attributes and file naming
  - Virtual variables (functions in CDAWlib, compute values on-the-fly)
      - »
- Checking and validation functions
  - Against ISTP/SPDF standards
  - For PRBEM, MMS or other specified project compliance reporting
- New JavaScript SKTeditor plans to add capability to add SPASE metadata at the same time when creating a dataset
  - Incorporate Lee Bargatze's ADAPT business logic to reduce effort

SKTeditor


ISTP CDF Skeleton Editor File Help
SKTEditor: timed.cdf
Information
ISTP Global Attributes
Variables
Space Physics Data Facility
NASA
Goddard Space Flight Center

For information on the SKTEditor see: http://spdf.gsfc.nasa.gov/skteditor
For information on the ISTP Guidelines see: http://spdf.gsfc.nasa.gov/sp_use_of_cdf.html
Program version: 1.3.1.31
Program build date: 2016/07/25 11:54:58 -0400

File Statistics

File Name	timed.cdf
CDF Library Version	3.6.2 1
CDF File Version	2.6.7
Library Last Leap Second	2015-07-01
File Last Leap Second	unknown
CDF Encoding	Network
CDF Majority	Row
Number of Variables	177
Number of Attributes	89 Attributes (60G/29V)
Compression	No compression
Compression Percent	100
Checksum	None

Ready

Required

Project: TIMED
Source / Spacecraft Name: TIMED
Descriptor / Instrument Name: GUVI>Global UltraViolet Imager
Data Type: L1CDisk
File Naming Convention: source_datatype_descriptor_ yyyyMMdd
PI Name: Andrew Christensen
PI Affiliation: Aerospace
Discipline: Space Physics>Ionospheric Science
Mission Group: TIMED
Instrument Types: Imaging and Remote Sens
Data Version: 1
Logical Source / Short Dataset Description: TIMED GUVI Level 1C data.
Extended Dataset Descriptive Text

Recommended

Acknowledgement
Rules of Use
ADID Ref
Time Resolution
Generated by
Generation Date
Link Text (describing on-line data)
Link Title
HTTP Link
Modification History

Ready

Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
Intensity1_rectified
Intensity2_rectified
Intensity3_rectified
Intensity4_rectified
Intensity5_rectified
Intns1_rec_MC_log_m
Intns2_rec_MC_log_m
Intns3_rec_MC_log_m
Intns4_rec_MC_log_m
Intns5_rec_MC_log_m
Intns1_rec_MC_lin_mc
Intns2_rec_MC_lin_mc
Intns3_rec_MC_lin_mc
Intns4_rec_MC_lin_mc
Intns5_rec_MC_lin_mc
Intns1_rec_NP_log

CDF Specifications

Name	Data Type	Time Varying	Dimensions	Compression
Intns1_rec_MC_	CDF_FLOAT/1	true	2:[1647,130]	No compression
	Sparse Recd		Pad Value	
	None		0.0	

Description

Expanded Label: H (1216 A) log10 Intensity Mercator
One-Line Description: [Mercator Projection by Orbit, Log10 Scaling] H Ly-alpha (1216 A) Intensities [NO LISTINGS]
Variable Notes

Axis Information

Label 1 Label 2 Label 3
Label 4 Label 5 Label 6
Scale Type
Format: e11.4
Units: Log10 of Int

Value Uncertainty
Plus Minus

Plot Information
Variable Type: Data
Display Type: Map Image
Display Arguments: thumbsize>250>map_proj>9>x=geographlat,y=geographlon

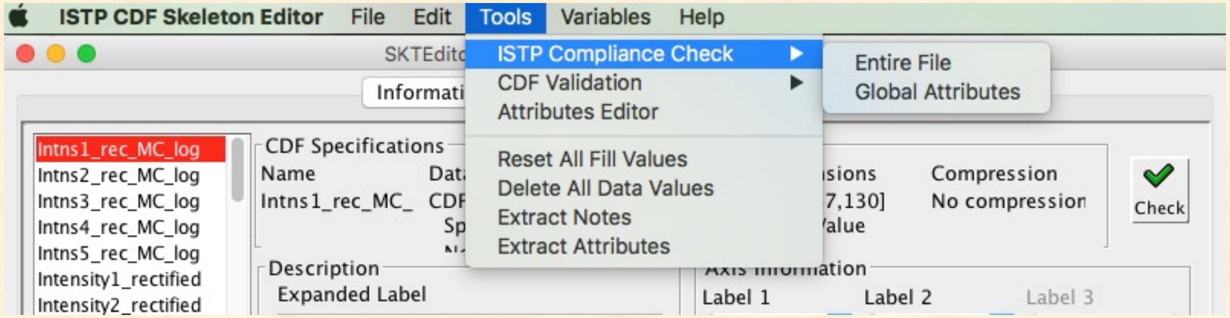
Depends
Depend 0: Epoch2
Depend 1: GeographLat
Depend 2

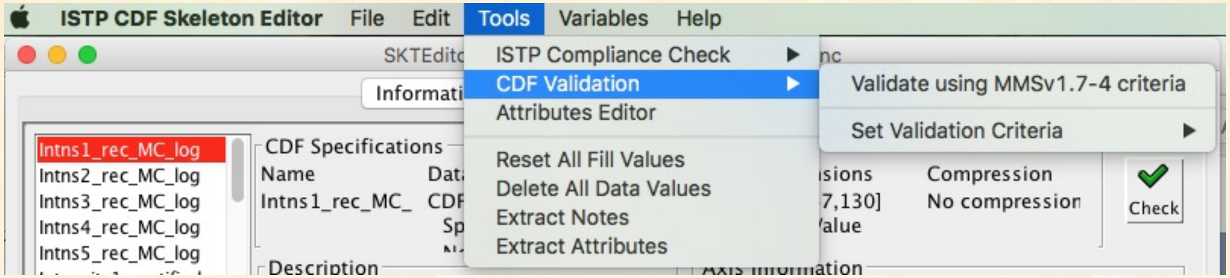
Virtu...
Fun... cor
Co...
F...
Inte...

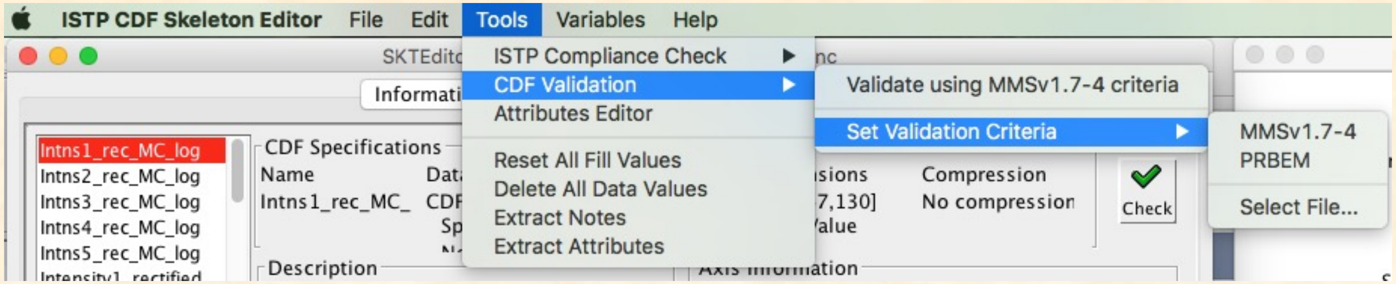
Valid Min: Fill all with selected value 1.0
Valid Max: Fill all with selected value 5.0

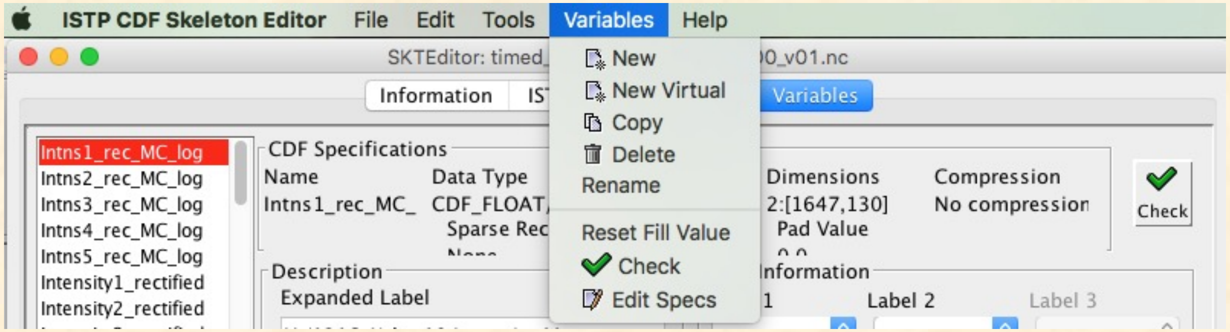
Ready

```
The following variables are not ISTP-compliant:
        Intns1_rec_MC_log
                DISPLAY_TYPE attribute value 'map_image>THUMBSIZE>250>MAP_PROJ>9>x=GeographLat,y=GeographLon' is not all lower case.
                DISPLAY_TYPE attribute value changed to 'map_image>thumbsize>250>map_proj>9>x=geographlat,y=geographlon'.
                DISPLAY_TYPE error: invalid keyword 'thumbsize'
                DEPEND_2 is not 1 dimensional
                DEPEND_2 is wrong size
                DEPEND_1 is not 1 dimensional
```

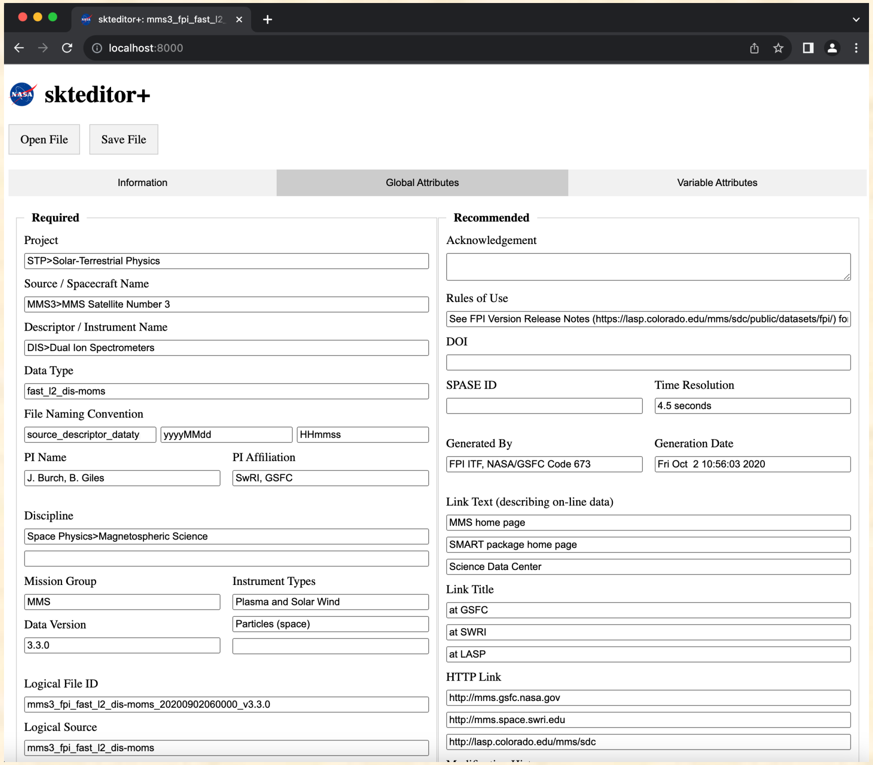

ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
SKTEdito
Informati
ISTP Compliance Check ▶
CDF Validation ▶
Attributes Editor
Reset All Fill Values
Delete All Data Values
Extract Notes
Extract Attributes
Entire File
Global Attributes
Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
Intensity1_rectified
Intensity2_rectified
CDF Specifications
Name
Dat
Intns1_rec_MC_
CDF
Sp
Description
Expanded Label
isions
7,130]
/alue
Compression
No compression
Check
Axis Information
Label 1
Label 2
Label 3


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
SKTEdito
nc
Informati
ISTP Compliance Check ▶
CDF Validation ▶
Attributes Editor
Reset All Fill Values
Delete All Data Values
Extract Notes
Extract Attributes
Validate using MMSv1.7-4 criteria
Set Validation Criteria ▶
Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
CDF Specifications
Name
Dat
Intns1_rec_MC_
CDF
Sp
Description
isions
7,130]
/alue
Compression
No compression
Check
Axis Information


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
SKTEdit
nc
Informati
ISTP Compliance Check ▶
CDF Validation ▶
Attributes Editor
Reset All Fill Values
Delete All Data Values
Extract Notes
Extract Attributes
Validate using MMSv1.7-4 criteria
Set Validation Criteria ▶
MMSv1.7-4
PRBEM
Select File...
Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
Intensity1_rectified
CDF Specifications
Name
Dat
Intns1_rec_MC_
CDF
Sp
Description
isions
7,130]
/alue
Compression
No compression
Check
Axis Information


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
SKTEditor: timed_
0_v01.nc
Information
IS
Variables
New
New Virtual
Copy
Delete
Rename
Reset Fill Value
Check
Edit Specs
Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
Intensity1_rectified
Intensity2_rectified
CDF Specifications
Name
Data Type
Intns1_rec_MC_
CDF_FLOAT
Sparse Rec
Description
Expanded Label
Dimensions
2:[1647,130]
Pad Value
Compression
No compression
Check
Information
Label 2
Label 3

Rewrite of SKTeditor in JavaScript for Laying Out Datasets and Adding ISTP and SPASE Metadata



Required

Project: STP>Solar-Terrestrial Physics

Source / Spacecraft Name: MMS3>MMS Satellite Number 3

Descriptor / Instrument Name: DIS>Dual Ion Spectrometers

Data Type: fast_l2_dis-moms

File Naming Convention: source_descriptor_dataty | yyyyMMdd | HHmmss

PI Name: J. Burch, B. Giles

PI Affiliation: SwRI, GSFC

Discipline: Space Physics>Magnetospheric Science

Mission Group: MMS

Instrument Types: Plasma and Solar Wind; Particles (space)

Data Version: 3.3.0

Logical File ID: mms3_fpi_fast_l2_dis-moms_20200902060000_v3.3.0

Logical Source: mms3_fpi_fast_l2_dis-moms

Recommended

Acknowledgement:

Rules of Use: See FPI Version Release Notes (https://lasp.colorado.edu/mms/sdc/public/datasets/fpi/) fo

DOI:

SPASE ID:

Time Resolution: 4.5 seconds

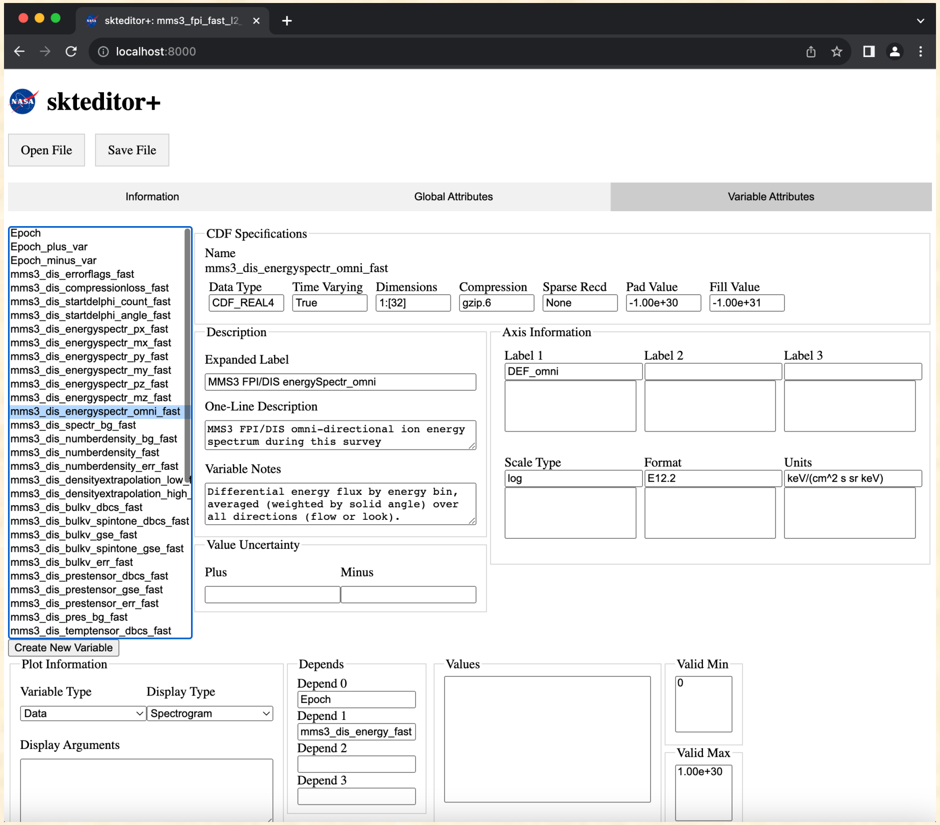
Generated By: FPI ITF, NASA/GSFC Code 673

Generation Date: Fri Oct 2 10:56:03 2020

Link Text (describing on-line data):
MMS home page
SMART package home page
Science Data Center

Link Title:
at GSFC
at SWRI
at LASP

HTTP Link:
http://mms.gsfc.nasa.gov
http://mms.space.swri.edu
http://lasp.colorado.edu/mms/sdc



Epoch
Epoch_plus_var
Epoch_minus_var
mms3_dis_errorflags_fast
mms3_dis_compressionloss_fast
mms3_dis_startdelphi_count_fast
mms3_dis_startdelphi_angle_fast
mms3_dis_energyspectr_px_fast
mms3_dis_energyspectr_mx_fast
mms3_dis_energyspectr_py_fast
mms3_dis_energyspectr_my_fast
mms3_dis_energyspectr_pz_fast
mms3_dis_energyspectr_mz_fast
mms3_dis_energyspectr_omni_fast
mms3_dis_spectr_bg_fast
mms3_dis_numberdensity_bg_fast
mms3_dis_numberdensity_fast
mms3_dis_numberdensity_err_fast
mms3_dis_densityextrapolation_low_
mms3_dis_densityextrapolation_high_
mms3_dis_bulkv_dbcs_fast
mms3_dis_bulkv_spintone_dbcs_fast
mms3_dis_bulkv_gse_fast
mms3_dis_bulkv_spintone_gse_fast
mms3_dis_bulkv_err_fast
mms3_dis_prestensor_dbcs_fast
mms3_dis_prestensor_gse_fast
mms3_dis_prestensor_err_fast
mms3_dis_pres_bg_fast
mms3_dis_temptensor_dbcs_fast

CDF Specifications

Name: mms3_dis_energyspectr_omni_fast

Data Type	Time Varying	Dimensions	Compression	Sparse Recd	Pad Value	Fill Value
CDF_REAL4	True	1:[32]	gzip.6	None	-1.00e+30	-1.00e+31

Description

Expanded Label: MMS3 FPI/DIS energySpectr_omni

One-Line Description: MMS3 FPI/DIS omni-directional ion energy spectrum during this survey

Variable Notes: Differential energy flux by energy bin, averaged (weighted by solid angle) over all directions (flow or look).

Axis Information

Label 1	Label 2	Label 3
DEF_omni		

Scale Type	Format	Units
log	E12.2	keV/(cm^2 s sr keV)

Value Uncertainty

Plus:

Minus:

Plot Information

Variable Type: Data

Display Type: Spectrogram

Display Arguments:

Depends

Depend 0: Epoch

Depend 1: mms3_dis_energy_fast

Depend 2:

Depend 3:

Values

Valid Min: 0

Valid Max: 1.00e+30

Next Steps

- Bobby to finish reviewing the existing documents at

  https://github.com/IHDE-Alliance/ISTP_metadata

  but looking for feedback on that draft

- Form user group or oversight committee to define changes

  Who would like to participate?

- Add content from **mission-specific documents** that reference the ISTP guidelines

- Add explanatory material from the CF Conventions (https://cfconventions.org/) that also apply in heliophysics

- Add crosswalk with SPASE metadata

- We are looking for feedback on whether this is a suitable path forward, and for feedback on its layout and content

Backup slides

Creating an ISTP/IACG CDF/NetCDF: Understand the Data to be Loaded

- What are the key data quantities
  - What is their definition/meaning?
  - How are they going to be named?
    - N.B. MMS parameter naming convention: scId_instrumentID_paramName
      - »
- Understand (at the dataset level)
  - Dimensionality and dependencies
  - Variance with time and dimension
    - ISTP/SPDF conventions allow >1 time variable in a file
    - Carry slowly-varying data as variables rather than in attributes
- General rule is to capture relationships in the structure
  - Otherwise capture relationships in variable attributes
  - Want relationships to be logically-structured and machine-readable
    - Available for more general-purpose codes to exploit
- Let CDF/NetCDF deal with mechanics of efficient data storage
  - Once more: lay out data by what's science logical and useful
    - E.g. methods to handle slowly-varying data include setting "sparse=sRecords.PREV" in CDFs

Upcoming Activities

- CDF
  - Ongoing maintenance, performance improvements
  - CDF beginner's guide
  - Python library: add WCS time conversions
  - Adapt NetCDF command line tools like NCO.sf.net for CDFs for operations on files

- ISTP/SPDF Guidelines
  - Will soon add SPASE and DOI global attributes to CDAWeb datasets via Master CDFs when available and expose in CDAWeb interface
  - Better document Guidelines on Github with mission-specific metadata as well, but want to keep flexible for interactions with missions and enabling framework for CDAWeb services

- Rewrite SKTeditor in JavaScript or similar and include SPASE fields
- Changes are driven by active archiving needs and new technology

Some Standards and Conventions

- **SPASE** *http://www.spase-group.org* dataset descriptions for easy searching
- **Heliophysics Data Portal** https://heliophysicsdata.sci.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*
  - Defining additional standard attributes: Cluster, THEMIS, RBSP (PRBEM), MMS, etc.
- **Dataset naming and file naming** recommendations *https://spdf.gsfc.nasa.gov/guidelines/filenaming_recommendations.html*

  and file naming templates:

  https://github.com/hapi-server/uri-templates/wiki/Specification **$Y/data_$Y_j_idx.cdf**
- **CDF** *https://cdf.gsfc.nasa.gov* scientific data format (including pure Python library *https://github.com/MAVENSDC/cdflib*)
  - Time variable types *https://cdf.gsfc.nasa.gov/html/leapseconds_requirements.html*
- **NetCDF** *https://www.unidata.ucar.edu/software/netcdf/*
- **FITS** *https://fits.gsfc.nasa.gov/*
- **UDunits** *www.unidata.ucar.edu/software/udunits/*
- Tools enabled by standards: CDAWeb and CDAWlib IDL/Python library, Autoplot http://autoplot.org, SPEDAS *http://spedas.org* IDL/Python library

Formats in NASA Space Science

- Standard formats
  - **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** added **CDF-A** as standard format (PDS-3, PDS-4, JPEG): CDF with ISTP/SPDF Guidelines and two SPASE attributes, but no compression or sparse variables

- ICON/GOLD metadata uses the ISTP/SPDF guidelines in NetCDFs, NetCDF4 Classic model with no groups or user-defined variable types, time is unlimited dimension

- SPDF has converters between CDF, CDFML, NetCDF, HDF, FITS, and to PDS-3

NetCDF Issues

- No predefined time variable types
  - Time not always the unlimited dimension
  - CDAWeb adds CDF_TIME_TT2000 virtual variables for NetCDF datasets, computed from various time schemes (base time, time units)
- CDAWeb adds missing Fillval, Validmin/max, Var_type, depend_0, and other attributes
- NetCDF to CDF converter adds attributes to store version, dimensions, sizes, compression, chunking, and string (not character) information
- Compression requires careful block size determination
- CDF to NetCDF converter converts time variables to binary or encoded string forms
- Supports only NetCDF4 Classic model with no groups or user-defined variable types