Standardized Heliophysics ISTP Metadata Guidelines and Other Guidelines for Dataset Production and Description

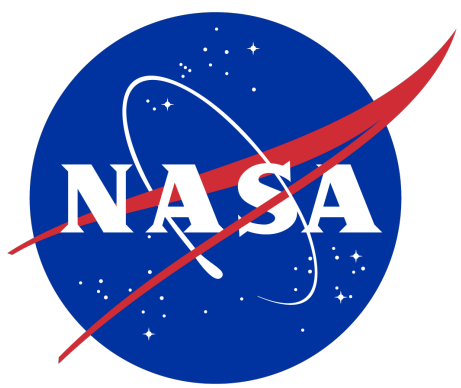
Robert Candey on behalf of SPDF Team

Robert M. Candey (NASA), Robert E. McGuire (NASA, retired), Tamara J. Kovalick (NASA/Adnet), Bernard T. Harris (NASA), Lan Jian (NASA), Andriy Koval (NASA/UMBC), Eric Grimes (NASA/Adnet)

Space Physics Data Facility (SPDF)

https://spdf.gsfc.nasa.gov

NASA Goddard Space Flight Center



Poster: IN31C-0668

Fall AGU 2023 Dec. 13

History

In the 1990s, NASA, the European Space Agency (ESA), the Institute of Space and Astronautical Science (ISAS) of Japan, and the Academy of Sciences (Russia)/Rosaviakosmos collaborated on the International Solar-Terrestrial Physics (ISTP) science initiative to combine resources and scientific communities for coordinated, simultaneous investigations of the Sun-Earth space environment over an extended period of time. ISTP was the flagship program for the space physics scientific community, the "great observatory" of the 1990's. As part of this effort, the community developed the ISTP Metadata Guidelines and other conventions for describing and naming the datasets. These guidelines were later adopted by the Interagency Consultative Group (IACG), and now by most heliophysics missions. Widely-used data display and analysis tools depend on the guidelines. Many missions have codified their use of the guidelines, and the International Heliophysics Data Environment Alliance (IHDEA.net) through the NASA Space Physics Data Facility (SPDF https://spdf.gsfc.nasa.gov) is extending and further standardizing the metadata conventions (https://github.com/IHDE-Alliance/ISTP_metadata).

https://spdf.gsfc.nasa.gov/pub/documents/SPDF/presentations/

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

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 metadata]
  - CDF in the rest of Heliophysics [ISTP Guidelines metadata]

- PDS added CDF-A as standard format (PDS-3, PDS-4, JPEG): CDF with ISTP Guidelines and two SPASE attributes, but no compression or sparse variables (SPDF can then host PDS CDF-A files easily with its extra services)

- ICON/GOLD metadata uses the ISTP 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

Dataset creation: 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 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

ISTP 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 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

- Guidelines in Markdown format at *https://github.com/IHDE-Alliance/ISTP_metadata*
- 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, 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-group.org 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
- Add additional conventions for record-varying DEPEND metadata to handle multi-dimensional dependent variables

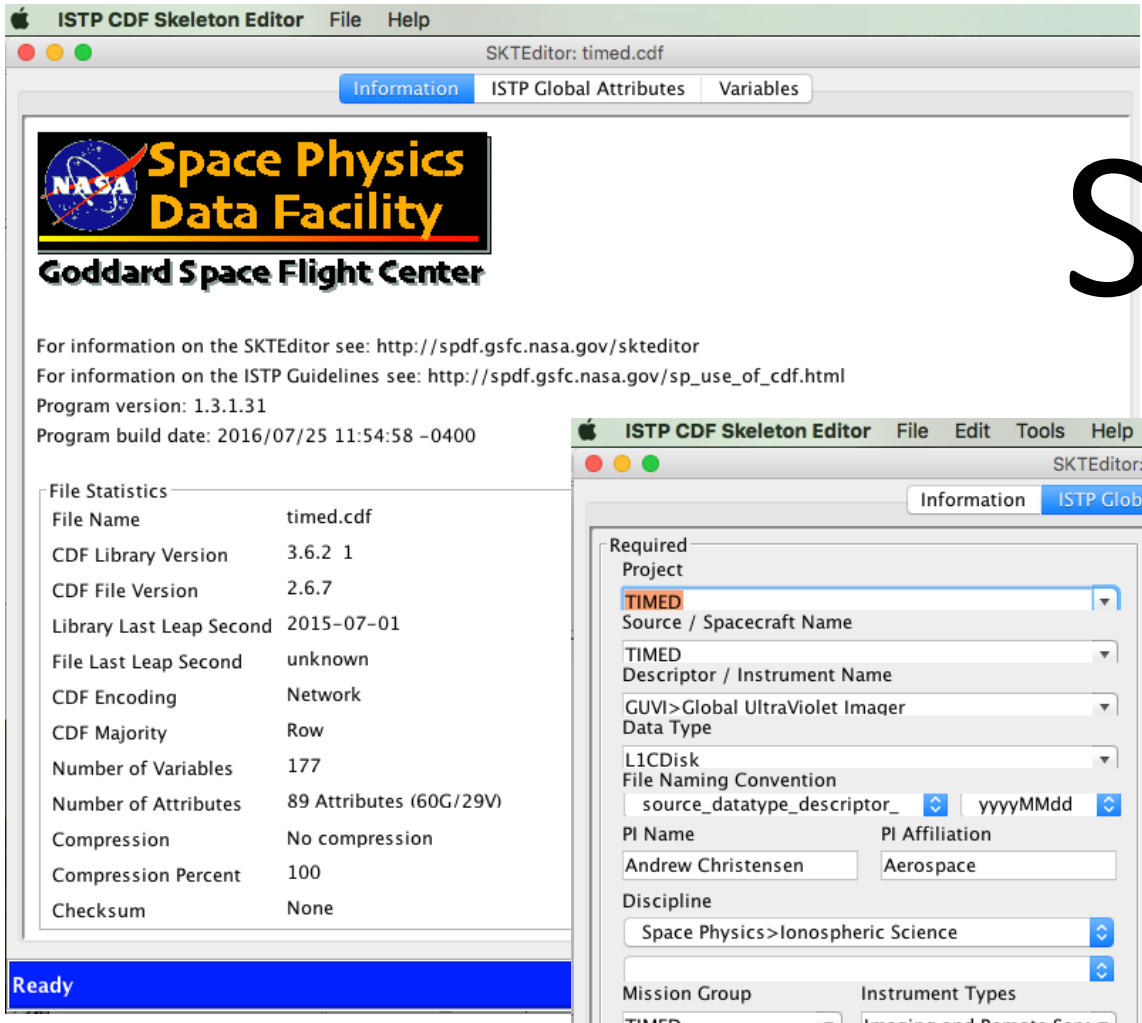
Tool to Create/Edit a CDF/NetCDF File Compliant to ISTP Standard

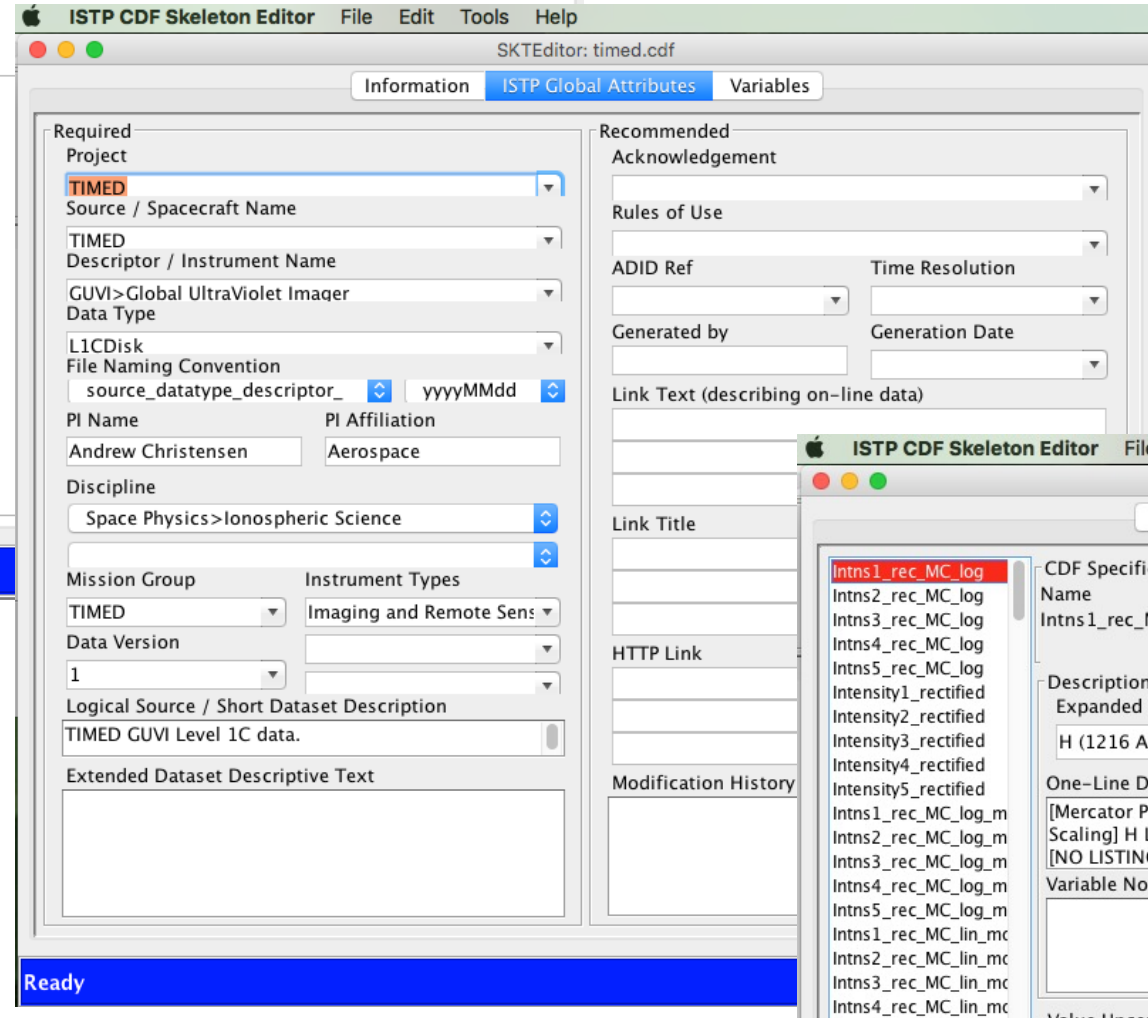
- SKTeditor is a Java, web-start application, soon to be in JavaScript
  - Guide designers to good choices consistent with ISTP guidelines
  - Create new CDF/NetCDF or check/correct then modify existing skeleton file
- Guided by the interface flow, add or edit
  - Scalar and higher-dimensional variables, 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)

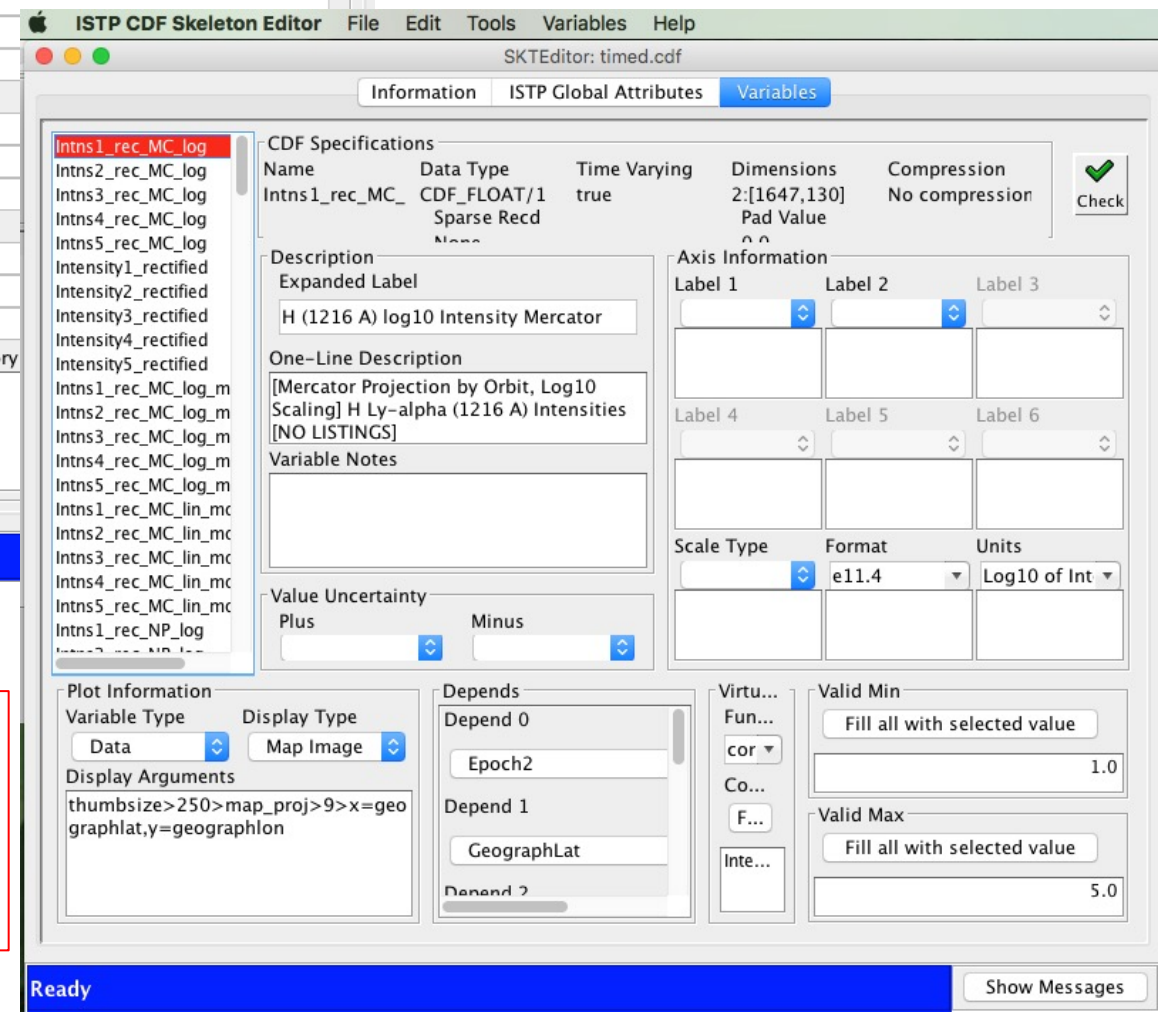
»

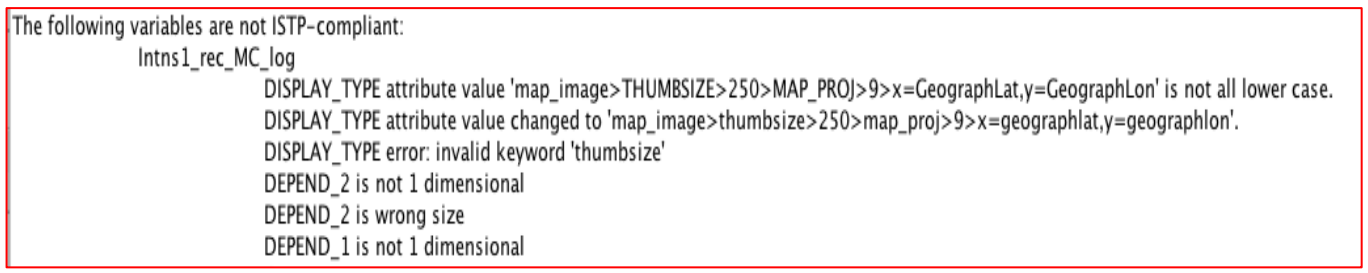
- Checking and validation functions
  - Against ISTP standards
  - For PRBEM, MMS or other specified project compliance reporting
- New JavaScript SKTeditor plans to add capability to add SPASE metadata at the same time when creating a dataset
  - Incorporate Lee Bargatze's ADAPT business logic to reduce effort

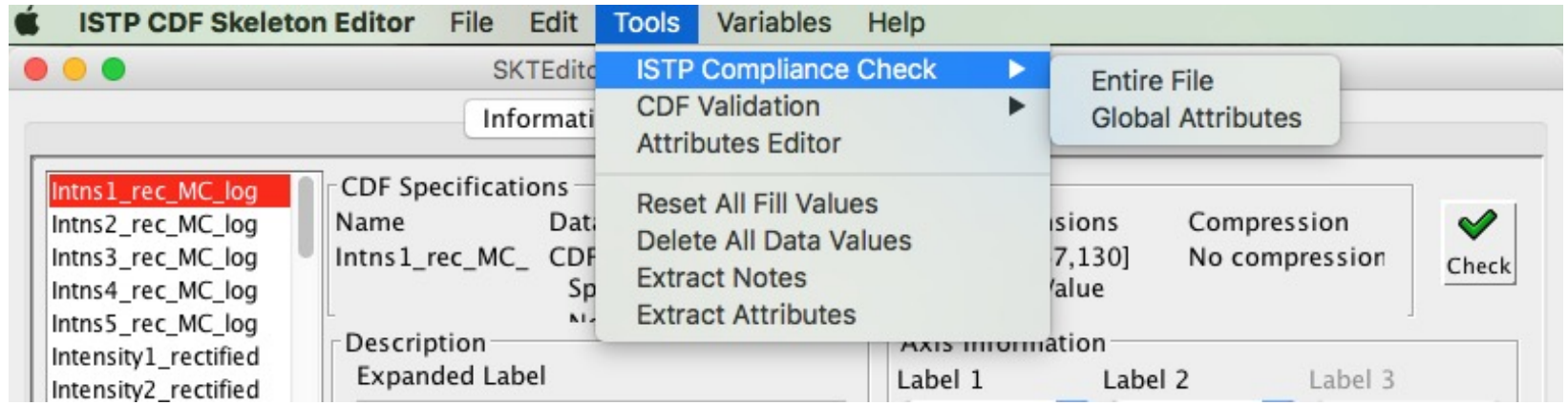
SKTeditor (Java)

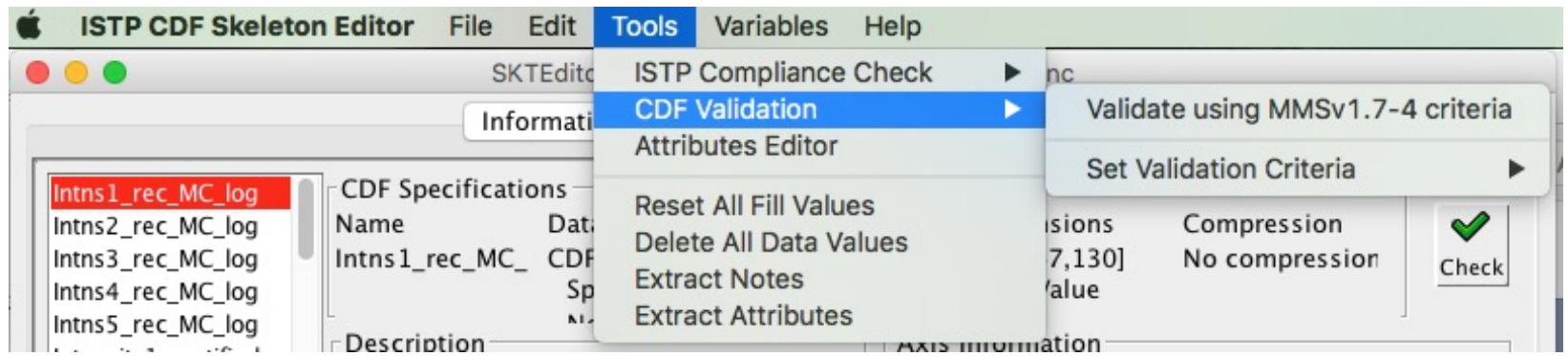


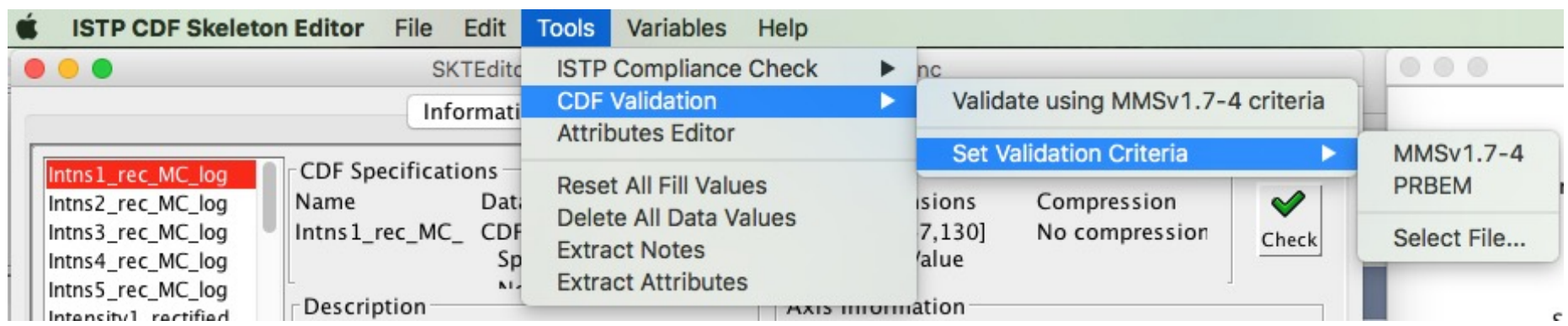


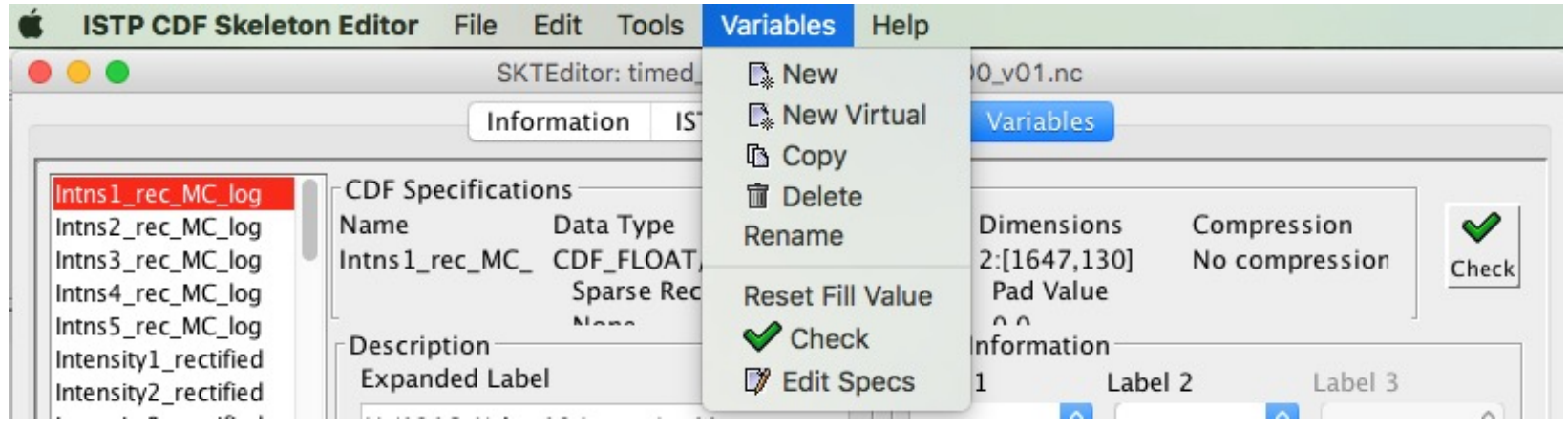




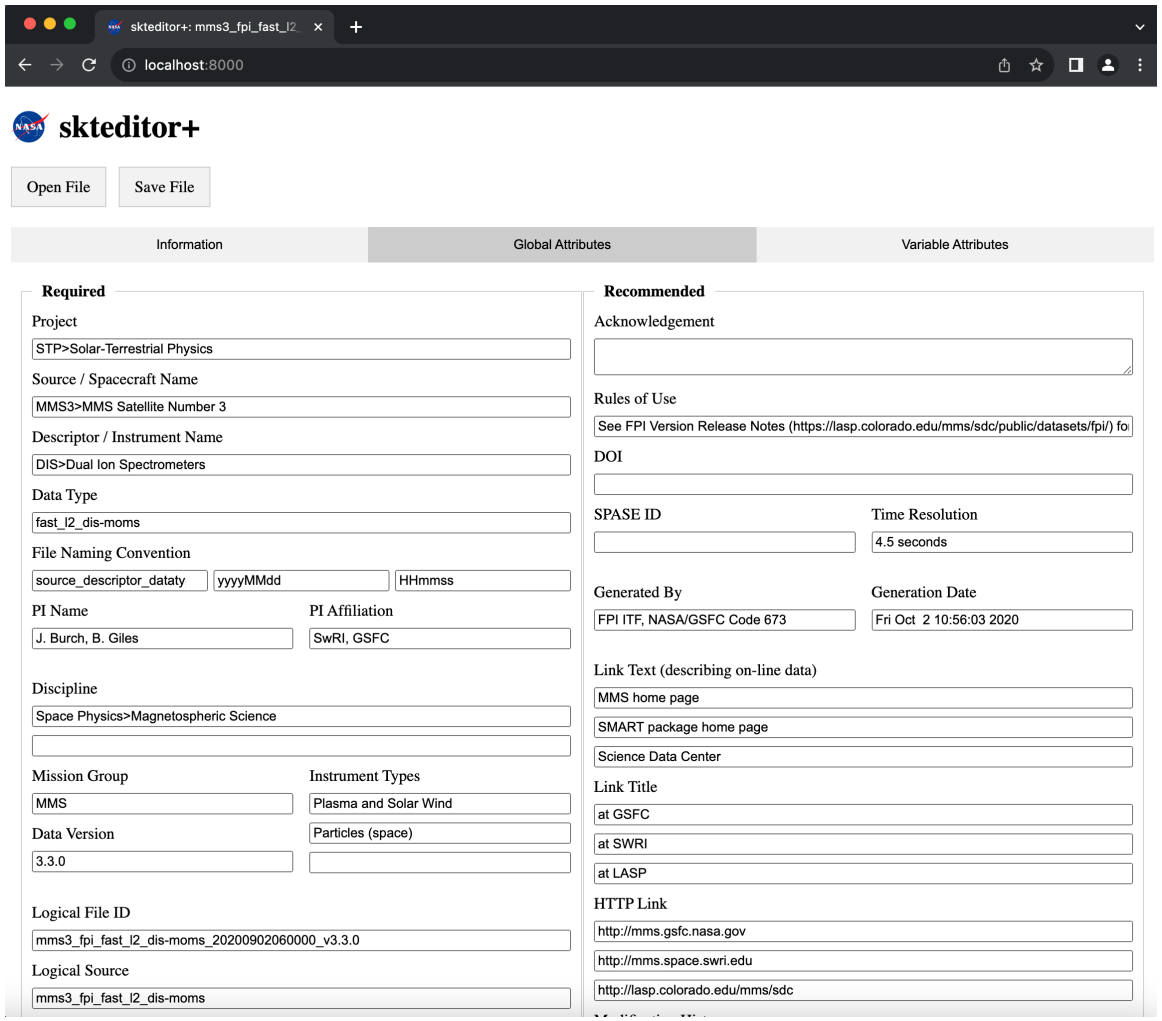

ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
ISTP Compliance Check
CDF Validation
Attributes Editor
Reset All Fill Values
Delete All Data Values
Extract Notes
Extract Attributes
Entire File
Global Attributes
Intns1_rec_MC_log
Intns2_rec_MC_log
Intns3_rec_MC_log
Intns4_rec_MC_log
Intns5_rec_MC_log
Intensity1_rectified
Intensity2_rectified
CDF Specifications
Name
Description
Expanded Label
Compression
No compression
Check
Label 1
Label 2
Label 3


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
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
Description
Compression
No compression
Check


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
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
CDF Specifications
Name
Description
Compression
No compression
Check


ISTP CDF Skeleton Editor
File
Edit
Tools
Variables
Help
New
New Virtual
Copy
Delete
Rename
Reset Fill Value
Check
Edit Specs
Information
Variables
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
Description
Expanded Label
Dimensions
Compression
No compression
Pad Value
Check
Label 2
Label 3

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



Next Steps for ISTP Metadata Guidelines

- Need 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

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 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

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 model with no groups or user-defined variable types