The Value of Data and Metadata Standardization for Interoperability in Giovanni
Or: Why your product's metadata causes us headaches!

IN41B-0039

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

Christine Smit[1,2], Mahabaleshwara Hegde[1,3], Richard Strub[1,3], Keith Bryant[1,3], Angela Li[1], Maksym Petrenko[1,3]

Abstract

Giovanni is a data exploration and visualization tool at the NASA Goddard Earth Sciences Data Information Services Center (GES DISC). It has been around in one form or another for more than 15 years. Giovanni calculates simple statistics and produces 22 different visualizations for more than 1600 geophysical parameters from more than 90 satellite and model products.

Giovanni relies on external data format standards to ensure interoperability, including the NetCDF CF Metadata Conventions. Unfortunately, these standards were insufficient to make Giovanni's internal data representation truly simple to use. Finding and working with dimensions can be convoluted with the CF Conventions. Furthermore, the CF Conventions are silent on machine-friendly descriptive metadata such as the parameter's source product and product version.

In order to simplify analyzing disparate earth science data parameters in a unified way, we developed Giovanni's internal standard. First, the format standardizes parameter dimensions and variables so they can be easily found. Second, the format adds all the machine-friendly metadata Giovanni needs to present our parameters to users in a consistent and clear manner. At a glance, users can grasp all the pertinent information about parameters both during parameter selection and after visualization.

https://giovanni.gsfc.nasa.gov/giovanni/

More about AIRS-only products!

"AIRS-only Product in Giovanni for Exploring Up-to-date AIRS Observation and Comparing with AIRS+AMSU Product"

Poster Session A11A
Calibration and Validation of Passive Satellite Earth Observations and Products

Finding dimensions associated with data variables

Analyzing data from multiple products presents unique challenges.
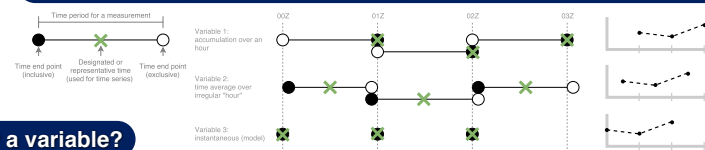
General issues:
- Where are the time / latitude / longitude / height / pressure associated with this data?
- What are the dimension variable bounds?
- Are the bounds inclusive or exclusive?

Special headaches with time:
- How do I convert the units into a more universal representation?
- How do I match data from different products in time so that I can compare them?
- If the data represents a day, do I really need to specify the time down to the last second?
- If I have to draw a single dot for each time step in a time series, where do I draw the dot?

Question #1: Where and when is this data? Please don't make this hard...

Yeah. And don't use different time units in every file of your product. That's just cruel.

TL;DR: CF is not enough...

What metadata is needed to fully qualify a variable?
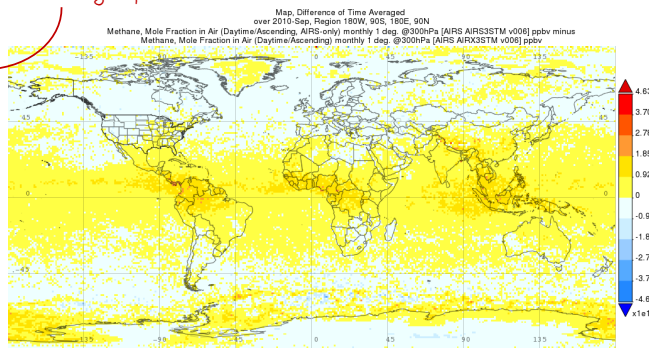
Descriptive name
something that tells you what this variable measures

Temporal resolution
the time between sequential measurements
Geographic resolution
how much area each data point represents
Platform/Instrument
which instrument collected the data
Product/Collection
what data product the variable came from
Version
what version of the algorithm was used to create this data

Units
the units of the data measurement

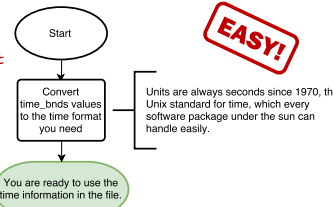*Why can't we describe this in CF? It's really important!*

Finding time in Giovanni

Start

Convert time_bnds values to the time format you need

Units are always seconds since 1970, the Unix standard for time, which every software package under the sun can handle easily.

You are ready to use the time information in the file.

EASY!

Same temporal resolution, very different time metadata

Time period for a measurement

Time end point (inclusive)
Designated or representative time (used for time series)
Time end point (exclusive)

Variable 1: accumulation over an hour
Variable 2: time average over irregular "hour"
Variable 3: instantaneous (model)

... and this just scratches the surface ... You would not believe the weird stuff we've seen!

Map, Difference of Time Averaged over 2010-Sep, Region 180W, 90S, 180E, 90N
Methane, Mole Fraction in Air (Daytime/Ascending, AIRS-only) monthly 1 deg. @300hPa [AIRS AIRS3STM v006] ppbv minus
Methane, Mole Fraction in Air (Daytime/Ascending, AIRS-only) monthly 1 deg. @300hPa [AIRS AIRX3STM v008] ppbv
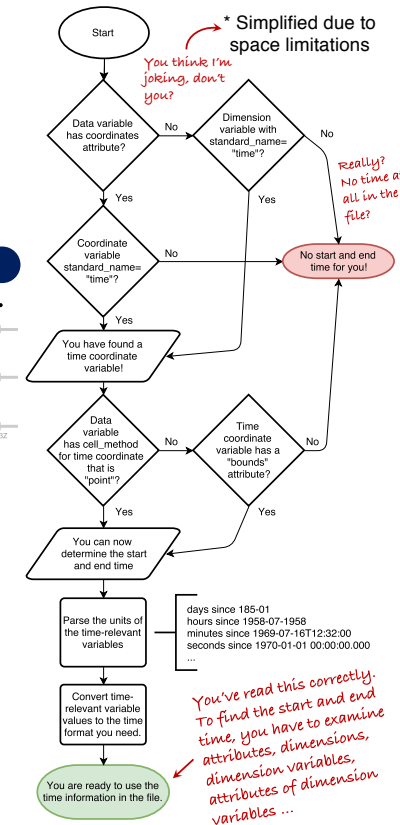
Methane, Mole Fraction in Air (Daytime/Ascending, AIRS-only) monthly 1 deg. @300hPa [AIRS AIRS3STM v006] ppbv

descriptive/long name | temporal resolution | geographic resolution | vertical layer and units | instrument | product | version | data units

Finding time using CF Conventions*

* Simplified due to space limitations

Start

Data variable has coordinates attribute?

Dimension variable with standard_name="time"?

You think I'm joking, don't you?

Really? No time at all in the file?

Coordinate variable standard_name="time"?

No start and end time for you!

You have found a time coordinate variable!

Data variable has cell_method for time coordinate that is "point"?

Time coordinate variable has a "bounds" attribute?

You can now determine the start and end time

Parse the units of the time-relevant variables

days since 185-01
hours since 1958-07-1958
minutes since 1969-07-16T12:32:00
seconds since 1970-01-01 00:00:00.000
...

Convert time-relevant variable values to the time format you need

You are ready to use the time information in the file.

You've read this correctly. To find the start and end time, you have to examine attributes, dimensions, dimension variables, dimension variables, attributes of dimension variables ...

Authors

[1] NASA Goddard Space Flight Center
[2] Telophase Corporation
[3] ADNET Systems Inc.