The Value of Data and Metadata Standardization for Interoperability in Giovanni

Or: Why your product’s metadata causes us headaches!

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 descriptive name, how much area each data point represents, what version of the algorithm was used to create this data, the time, and the units of the data measurement. This makes it hard to find and compare data from different products in time so that I can compare them.

TLDR: CF is not enough... You think I’m joking, don’t you?

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?

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

Finding time in Giovanni

Question #2: 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 onel.

Same temporal resolution, very different time metadata

Finding time using CF Conventions*

Data variable has coordinate attribute Yes

Can you find a time coordinate attribute? Yes

Who’s the start of this time coordinate? Yes

Where are the time / latitude / longitude / height / pressure associated with this data? No

Are the bounds inclusive or exclusive? No

You found a time coordinate attribute! Yes

Units are always seconds since 1970, the Unix standard for time, which every human can handle easily.

TLDR: CF is not enough... You think I’m joking, don’t you?

Finding time using CF Conventions

Data variable has coordinate attribute Yes

Can you find a time coordinate attribute? Yes

Who’s the start of this time coordinate? Yes

Where are the time / latitude / longitude / height / pressure associated with this data? No

Are the bounds inclusive or exclusive? No

You found a time coordinate attribute! Yes

Units are always seconds since 1970, the Unix standard for time, which every human can handle easily.

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