CF2-Group: Draft Climate/Forecast Conventions for Hierarchical Data+Metadata

Charlie Zender
Departments of Earth System Science and of Computer Science
University of California, Irvine

DIWG Telecon + EarthCube CF2
September 20, 2017

Seminar on Web

Use Cases of Groups

1. Collections ("suitcases") of data
2. Ensembles (one group per realization/station)
3. Discrete Sampling Geometries (groups instead of instance or "station" dimension)
4. Remote sensing
CF2-Group: Design Principles

1. Backwards compatible with CF1 ("CF-compliant" within each group)
2.Convertible to/from netCDF3 ("flattening", "dismembering", "inflating")
3. Amenable and robust to subsetting and aggregation: Use hierarchical principles of inheritance, scope
1. Newer netCDF4 atomic types are first-class citizens. Avoid netCDF4 "non-atomic" types like enum, vlen, compound, opaque, and user-defined types when compliance with CF1 is paramount, otherwise procedures to convert CF2-Group files to CF1 flat files may lose information or fail completely.
2. Attributes normally attached to variables should not be stored as Group Attributes (e.g., valid_min, units, _FillValue) even when the attribute values are identical across the group. This is not interoperable.
3. Group names should have no machine-readable relevance. Automatically generated group names (e.g., ensembles) are fine so long as their information content (e.g., realization number, station number) is separately stored as a Group Attribute.
4. Moving a self-contained group or branch of groups to another location, should not impact the interpretation of data. Named objects (dimensions, coordinates, variables) resolve to the nearest in-scope object of that name. Absolute or relative pathnames (containing "/") in metadata are discouraged, as they are fragile.

***Current Satellite Swath Proposal recommends always using full paths to locate "ancillary" information, e.g., coordinates="/geolocation/grid/lat /geolocation/grid/lon"
Current "Typical" NASA L2 Dataset
Nearest Coord. in Dimension's Scope

/ 

g1  g2  g3  g4  

g1g1  g1g2  g2g2  g3g1  g3g2  

g1g1g1  g2g2g1  g2g2g2  g3g1g1  g3g1g2
CF2-Group: Best Practices Options for locating out-of-group (meta)data

1. Scoping and inheritance always resolve
2. Absolute pathnames always
3. Relative pathnames always
4. Absolute or relative always
5. Scoping preferred and Absolute/relative allowed
6. Others?
Mapping CF2-Group to CF1 Flat Files

1. Flattening
   a. No namespace conflicts? Easy (Hyrax, NCO)
   b. Conflicts require algorithmic renaming
      i. Hyrax replaces path separator '/' with underscores when flattening
      ii. NCO can only "dismember" conflicts

2. Inflation
   a. Partial solutions only ("chicken and egg")
   b. NCO ncecat with Group Path Editing
   c. CF2 keys could facilitate conversion?
Supplementary Slides