Automating Metadata Compliance Checking

Oliver Chang
University of Miami, Miami, Florida
August 20, 2014

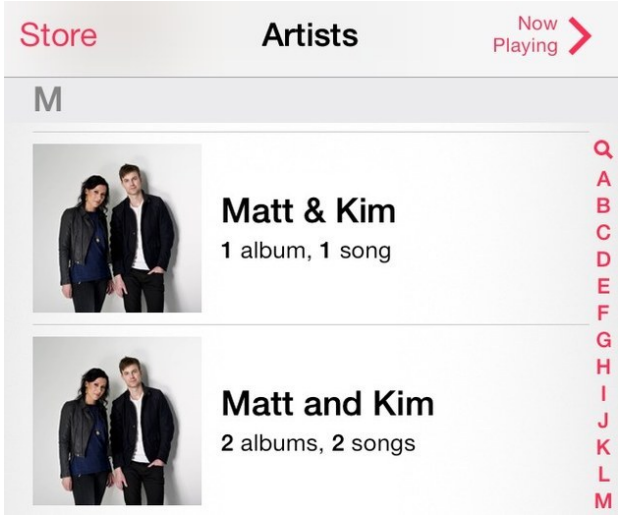
Background

* Check earth science data for valuable metadata, “data about data”
* Examples: units, date, time, author

* Incoming files vary in quantity and quality
* Need a tool for checking conformance
* Target three tests
  * Attribute Conventions for Dataset Discovery (ACDD)
  * Climate and Forecast (CF) Metadata Standards
  * GHRSST Data Specification, Version 2 (GDS2)
* Several existing tools already, run in terminal and as websites

Metadata Failures

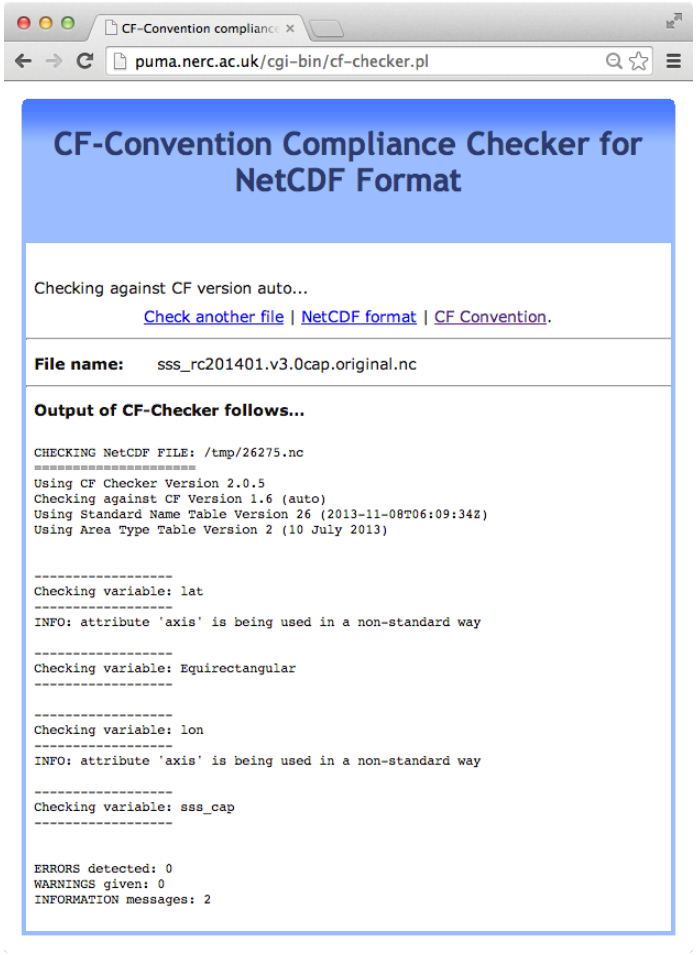
```
--- datasets/netcdf » ncdump -h GRCTellus.CSR.200208_201404.OCN.RL05.DSTvDPC1401.nc.nc
netcdf GRCTellus.CSR.200208_201404.OCN.RL05.DSTvDPC1401.nc {
dimensions:
        lon = 360 ;
        lat = 180 ;
        time = 131 ;
        bounds = 2 ;
variables:
        float lon(lon) ;
                lon:units = "degrees_east" ;
                lon:point_spacing = "even" ;
                lon:long_name = "Longitude" ;
        float lat(lat) ;
                lat:units = "degrees_north" ;
                lat:point_spacing = "even" ;
                lat:long_name = "Latitude" ;
        float time(time) ;
                time:units = "days since 2002-01-01 00:00:00" ;
                time:bounds = "time_bounds" ;
                time:long_name = "Time" ;
                time:calendar = "gregorian" ;
        float time_bounds(time, bounds) ;
                time_bounds:units = "Days since 2002-01-01 00:00:00" ;
                time_bounds:long_name = "time_bounds" ;
                time_bounds:calendar = "gregorian" ;
        float lwe_thickness(time, lat, lon) ;
                lwe_thickness:units = "cm" ;
                lwe_thickness:_FillValue = -9999.f ;
                lwe_thickness:long_name = "Liquid_Water_Water_Thickness" ;

// global attributes:
                :NC_GLOBAL.Conventions = "CF-1.6" ;
                :NC_GLOBAL.filename = "GRCTellus.CSR.200208_201404.OCN.RL05.DSTvDPC1401.nc" ;
                :NC_GLOBAL.institution = "JPL / GRACE-TELLUS" ;
                :NC_GLOBAL.variable = "water thickness" ;
                :NC_GLOBAL.unit = "cm_equiv_H2O" ;
                :NC_GLOBAL.platform = "GRACE" ;
                :NC_GLOBAL.sensor = "GRACE" ;
                :NC_GLOBAL.time_mean_removed = "2005.000 to 2010.999 (1/2005 to 12/2010)" ;
                :NC_GLOBAL.data_source = "Don P. Chambers" ;
                :NC_GLOBAL.data_source_version = "vDPC1401" ;
                :NC_GLOBAL.Longitudes = " LON1_NLONS_DLON=0.5. 360 1." ;
                :NC_GLOBAL.Latitudes = " LAT1_NLATS_DLAT=-89.5 180 1." ;
                :NC_GLOBAL.time_start = "200208" ;
                :NC_GLOBAL.time_end = "201404" ;
                :NC_GLOBAL.time_unit = "month-day-year" ;
                :NC_GLOBAL.postprocess1 = "DESTRIPED" ;
                :NC_GLOBAL.postprocess2 = "OCEAN_ATMOSPHERE_DEALIAS_MODEL (GAD), MONTHLY_AVE, ADDED BACK TO
                :NC_GLOBAL.postprocess3 = "GLOBAL MEAN OCEAN BOTTOM PRESSURE REMOVED" ;
                :NC_GLOBAL.filter = " gaussian" ;
                :NC_GLOBAL.Filter_Width_KM = "   500" ;
                :NC_GLOBAL.Filter_Max_Degree = "  40" ;
                :NC_GLOBAL.GIA_removed = "Paulson, Zhong, and Wahr, 2007, Geophys. J. Intl 171, 497-508, as
                :NC_GLOBAL.Citation = "Chambers, D. P. and Bonin, J. A.: Evaluation of Release-05 GRACE time
                :NC_GLOBAL.label = "GRC" ;
                :NC_GLOBAL.Mask = "OCEAN ONLY PIXELS (Sean Swanson ETOPO5 DEM-2013-10-15-ss_landmask-360-180
                :NC_GLOBAL.DATE_CREATED = "2014Jul11" ;
                :NC_GLOBAL.INPUT_FILENAME = "131_/acc1/vzraid2/vz11/akh/DATA/DESTRIPED_GRIDS/RL05/Chambers/2
}
```



No strictly enforced standard

Quantity of data makes it tedious, cumbersome to check manually

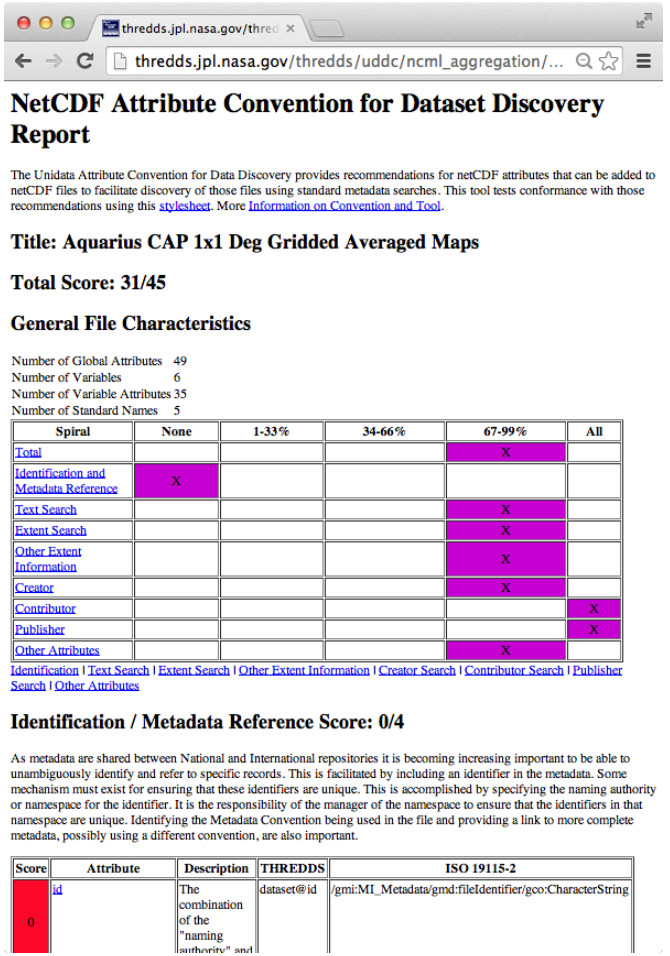
Example: CF Checker



* Generates # of errors, warnings, information results
* Web interface, thin wrapper around a Python script
* Not very detailed descriptions of tests
  * No info on what passed, what tests are available

Example: THREDDS UDDC

* "Spirals" = categories
* Very detailed descriptions
* Easy to parse visually
* Tightly integrated with other tools in THREDDS
* Infers some geospatial metadata fields



thredds.jpl.nasa.gov/thred ×

thredds.jpl.nasa.gov/thredds/uddc/ncml_aggregation/...

NetCDF Attribute Convention for Dataset Discovery Report

The Unidata Attribute Convention for Data Discovery provides recommendations for netCDF attributes that can be added to netCDF files to facilitate discovery of those files using standard metadata searches. This tool tests conformance with those recommendations using this stylesheet. More Information on Convention and Tool.

Title: Aquarius CAP 1x1 Deg Gridded Averaged Maps

Total Score: 31/45

General File Characteristics

Number of Global Attributes 49
Number of Variables 6
Number of Variable Attributes 35
Number of Standard Names 5

Spiral	None	1-33%	34-66%	67-99%	All
Total				X	
Identification and Metadata Reference	X				
Text Search				X	
Extent Search				X	
Other Extent Information				X	
Creator				X	
Contributor					X
Publisher					X
Other Attributes				X	

Identification | Text Search | Extent Search | Other Extent Information | Creator Search | Contributor Search | Publisher Search | Other Attributes

Identification / Metadata Reference Score: 0/4

As metadata are shared between National and International repositories it is becoming increasing important to be able to unambiguously identify and refer to specific records. This is facilitated by including an identifier in the metadata. Some mechanism must exist for ensuring that these identifiers are unique. This is accomplished by specifying the naming authority or namespace for the identifier. It is the responsibility of the manager of the namespace to ensure that the identifiers in that namespace are unique. Identifying the Metadata Convention being used in the file and providing a link to more complete metadata, possibly using a different convention, are also important.

Score	Attribute	Description	THREDDS	ISO 19115-2
0	id	The combination of the "naming	dataset@id	/gmi:MI_Metadata/gmd:fileIdentifier/gco:CharacterString

Example: GDS2 Validator

* Command line script
* Check global attributes, variables, and variable attributes
* Check existence and type
* A lot of text, hard to parse

```
Terminal
$ ./ghrsst_format_check.py -f ~/podaac/datasets/netcdf/20140508-MODIS_A-JPL-L2P-A
2014128024500.L2_LAC_GHRSST_N-v01.nc
 ---- Validate metadata and structure of a GHRSST GDS v2 file ----
                          ver 1.1


 Checking global attributes . . .
       Notice: Global attribute name GDS_version_id not recognized
       Notice: Global attribute name DSD_entry_id not recognized
       Notice: Global attribute name stop_date not recognized
       Notice: Global attribute name creation_date not recognized
       Notice: Global attribute name file_quality_index not recognized
       Notice: Global attribute name contact not recognized
       Notice: Global attribute name start_date not recognized

       Fatal: Required attribute  geospatial_lat_units was not found
       Fatal: Required attribute  geospatial_lon_units was not found
       Fatal: Required attribute  Metadata_Conventions was not found
       Fatal: Required attribute  keywords was not found
       Fatal: Required attribute  publisher_name was not found
       Fatal: Required attribute  id was not found
       Fatal: Required attribute  naming_authority was not found
       Fatal: Required attribute  uuid was not found
       Fatal: Required attribute  source was not found
       Fatal: Required attribute  standard_name_vocabulary was not found
       Fatal: Required attribute  creator_email was not found
       Fatal: Required attribute  publisher_url was not found
       Fatal: Required attribute  processing_level was not found
       Fatal: Required attribute  gds_version_id was not found
       Fatal: Required attribute  publisher_email was not found
       Fatal: Required attribute  keywords_vocabulary was not found
       Fatal: Required attribute  geospatial_lat_resolution was not found
       Fatal: Required attribute  time_coverage_start was not found
       Fatal: Required attribute  metadata_link was not found
       Fatal: Required attribute  date_created was not found
       Fatal: Required attribute  acknowledgment was not found
       Fatal: Required attribute  geospatial_lon_resolution was not found
       Fatal: Required attribute  license was not found
       Fatal: Required attribute  creator_name was not found
       Fatal: Required attribute  time_coverage_end was not found
       Fatal: Required attribute  summary was not found
       Fatal: Required attribute  project was not found
       Fatal: Required attribute  cdm_data_type was not found
       Fatal: Required attribute  file_quality_level was not found
       Fatal: Required attribute  creator_url was not found
        Review errors above!
```

Example: IOOS Compliance Checker

```
Terminal
$ ./cchecker.py -t=cf ~/podaac/datasets/netcdf/zos_AVISO_L4_199210-201012.nc
Running Compliance Checker on the dataset from: /Users/ochang/podaac/datasets/netcdf/zos_AVISO_L4_199
210-201012.nc

------------------------------------------------------------------------------
                The dataset scored 60 out of 69 points
                        during the cf check
------------------------------------------------------------------------------
                            Scoring Breakdown:


                              High Priority
------------------------------------------------------------------------------
    Name                            :Priority: Score
Variable names                          :3:     7/7
axis                                    :3:     9/9
convention_attrs                        :3:     2/2
conventions                             :3:     0/1
data_types                              :3:     7/7
dimension_names                         :3:     7/7
latitude                                :3:     4/4
longitude                               :3:     4/4
std_name                                :3:     4/4
time                                    :3:     4/4
units                                   :3:     4/4


                            Medium Priority
------------------------------------------------------------------------------
    Name                            :Priority: Score
all_features_are_same_type              :2:     0/0
contiguous_ragged_array                 :2:     0/0
coordinate_type                         :2:     3/3
coordinates_and_metadata                :2:     0/0
feature_type                            :2:     0/0
incomplete_multidim_array               :2:     0/0
indexed_ragged_array                    :2:     0/0
missing_data                            :2:     0/0
orthogonal_multidim_array               :2:     0/0
var                                     :2:     5/13


------------------------------------------------------------------------------
                Reasoning for the failed tests given below:

Name                            Priority:     Score:Reasoning
------------------------------------------------------------------------------
conventions                             :3:     0/ 1 : Conventions field is not
                                                       "CF-1.6"
var                                     :2:     5/13 :
    lat                                 :2:     1/ 2 :
        check_independent_axis_dimensio:2:     0/ 1 : The lat dimension for the
                                                       variable lat does not have
                                                       an associated coordinate
```

* Fully featured: summary scoring, hierarchy, detailed messages
* Checkers: CF, ACDD, IOOS Asset Concept
* Actively developed
* Linked to difficult to target dependencies, tied to a terminal output

Design

* Design inspiration from the IOOS tool, basically a thin wrapper around that tool
* Use the best part of the previous examples
* Target detailed descriptions, web interface, easy to add and update
* Many different interfaces (e.g. html, api, command line)

Development

* Rewritten ACDD and GDS2 checker tools
  * Take advantage of sorting, rich textual descriptions
  * Share much of the same code between the two checkers
* CF portion uses the IOOS compliance checker as a black box: feed it inputs, let it compute, and then inspect the results
  * Checks not only metadata but also compares its validity against the actual data – thorough but not very fast
* Data-driven design, tests are designed in a vaguely tree-like structure with a hierarchy of Python dicts
  * Easy to programmatically add new tests from configuration files

Demo