Improving Compliance for Earth Science Data Records

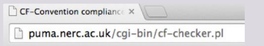
Oliver Chang[2], Edward.M.Armstrong@jpl.nasa.gov[1], David Foster[3]
NASA Jet Propulsion Laboratory/California Institute of Technology, Pasadena, California, USA[1]
University of Miami, Miami, Florida, USA[2]
Applied Science Associates, Rhode Island, USA[3]

2014 Fall AGU
IN31A-3713 -- ESDR Challenges

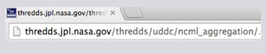
Introduction

One of the recurring challenges of creating earth science data records is to ensure a consistent level of metadata compliance at the granule level where important details of contents, provenance, producer, and data references are necessary to obtain a sufficient level of understanding. These details are important not just for individual data consumers but also for autonomous software systems. Two of the most popular metadata standards at the granule level are the Climate and Forecast (CF) Metadata Conventions and the Attribute Conventions for Dataset Discovery (ACDD). Many data producers have implemented one or both of these models including the Group for High Resolution Sea Surface Temperature (GHRSST) for their global SST products and the Ocean Biology Processing Group for NASA ocean color and SST products. While both the CF and ACDD models contain various level of metadata richness, the actual "required" attributes are quite small in number. Metadata at the granule level becomes much more useful when recommended or optional attributes are implemented that document spatial and temporal ranges, lineage and provenance, sources, keywords, and references etc. In this presentation we report on a new open source tool to check the compliance of netCDF and HDF5 granules to the CF and ACCD metadata models. The tool, written in Python, was originally implemented to support metadata compliance for netCDF records as part of the NOAA's Integrated Ocean Observing System. It outputs standardized scoring for metadata compliance for both CF and ACDD, produces an objective summary weight, and can be implemented for remote records via OPeNDAP calls. Originally a command-line tool, we have extended it to provide a user-friendly web interface. Reports on metadata testing are grouped in hierarchies that make it easier to track flaws and inconsistencies in the record. We have also extended it to support explicit metadata structures and semantic syntax for the GHRSST project that can be easily adapted to other satellite missions as well. Overall, we hope this tool will provide the community with a useful mechanism to improve metadata quality and consistency at the granule level by providing objective scoring and assessment, as well as encourage data producers to improve metadata quality and quantity.

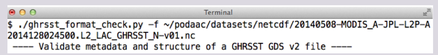
Other existing metadata checkers



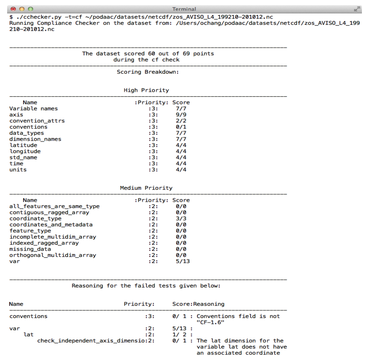
CF Project compliance checker
- Generates # of errors, warnings, information results
- Web interface, thin wrapper around a Python script
- Not very detailed descriptions of tests. No info on what passed, what tests are available



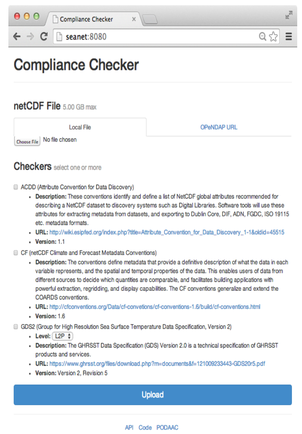
UDDC (THREDDS) compliance checker
- Spirals of categories
- Detailed output descriptions. Easy to parse visually
- Tightly integrated to THREDDS. Can be used by applying style sheet to NCML output (netCDF XML)



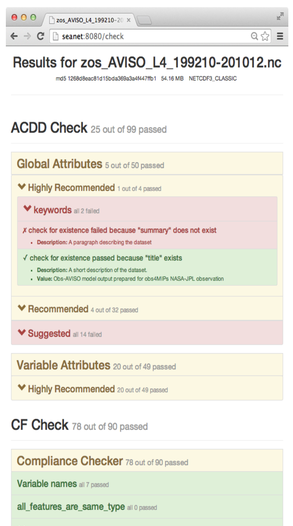
GHRSST compliance checker
- Command line tools
- Just test metadata existence and type
- Lots of text. Hard to parse.



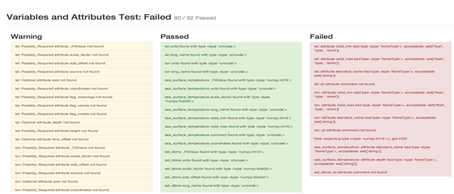
Example of IOOS tool command line output developed by the the NOAA IOOS. ASCII output summarizes the testing results and quantifies the level of metadata compliance to CF and ACDD metadata models



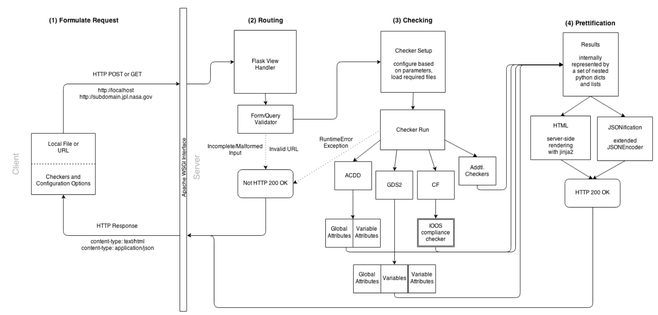
Web interface created by the PO.DAAC. A menu for CF, ACDD and GHRSST metadata checks is available. Tools requires local access to the file to be tested or remote access via OPeNDAP.



Example of web output. Tests are organize by metadata group (ACDD.CF/GHRSST), grouped in hierarchies and color coded for pass/fail status. Groups can be expanded. Testing results are quantified.



Yellow indicate partial passes. Green indicates all checks passed. Reds are failure.



System architecture and workflow of the compliance checker. The PO.DAAC has implemented a web based front end. Repacked the IOOS code for ACDD metadata checks to return attribute definition information. CF component is a blackbox.

In addition to the web frontend, there is a queryable API frontend that one can use to perform validation. It returns a JSON response that is identical to the data that's used to generate the HTML templates. The example shows the response to a command line call using UNIX curl command.

GET Example

```
$ curl http://seanet:8080/check?checkers=acdd&url-upload=https://github.com/ioos/compliance-checker/raw/master/compliance_checker/tests/data/2dim-grid.nc
{
  "fn": "2dim-grid.nc",
  "md5": "e7b1b61e7233ebad8a71140944105df",
  "models": "NETCDF4",
  "results": [
    {
      "description": "These conventions identify and define a list of NetCDF global attributes recommended for describing a NetCDF dataset to discovery systems such as Digital Libraries. Software tools will use these attributes for extracting metadata from datasets, and exporting to Dublin Core, DIF, ADN, FGDC, ISO 19115 etc. metadata formats.",
      "hash": 2524994885152638950,
      "name": "Attribute Convention for Data Discovery",
      "passed": 18,
      "results": [
        {
          "hash": 4800022785577525635,
          "name": "Global Attributes",
          "passed": 0,
          "results": [
            ...
```

Future Work and Implementation

- Public accessibility via PO. DAAC "Labs"
  - Initially only remote OPeNDAP granule testing
- Update code to handle ACDD 1.3x
- Update code to handle HDF5/netCDF4 enhanced group hierarchy data models
- Update code to modularize and extend CF testing (currently blackbox)
- Code maintained at JPL's software repository: code.jpl.nasa.gov