CMIP6 ESGF Publication Requirements

A WIP white paper; lead: Martin Juckes

Executive Summary

This document summarises the file syntax requirements which will be checked and verified prior to publication. Publication should not proceed when files do not comply.

Introduction

These requirements are a subset of the full specification. It is recognised that the management of the data generation and publication work-flow is extremely challenging, so this subset has been deliberately limited to the most important aspects of the file syntax which are easily verified and important for the smooth functioning of the ESGF distributed archive.

It is highly recommended that data providers run a more complete set of checks (software will be made available) and correct as much as possible before publication.

The use of the CMOR package to generate files consistent with the specification is highly recommended. The full specification is considerably more complex than the list below: this list should not be considered as a guide to the specifications.

The full details of the specifications in technical documents:
FGA: Filenames and Global Attributes for CMIP6 (in prep)
DRQ: CMIP6 Data Request (in prep).

1. File Name

The file name must match the structure specified in FGA. Components of the file name must match the appropriate global attributes, the variable name. The time range component of the file name (absent in fixed fields) must be consistent with the FGA specifications in terms of the number of characters and consistent with the temporal range in the file (the file name range must include the range of values in the file, with a margin which is less than the time interval implied by the declared frequency).

Frequency	**Date Syntax**	**Example Usage**
yr	YYYY	1986-2001
mon	YYYYMM	198001-198912
monClim	YYYYMM	198001-198912-clim

Day	YYYYMMDD	18600101-18691231
6hr, 3hr, 1hr, subhr	YYYYMMDDHHMM	198001010000-198101010000
1hrClimMon	YYYYMMDDHHMM	198001010000-198101010000-clim

2. Data Reference Syntax Global Attributes

The following global attributes, which are included in the Data Reference Syntax (DRS), must be present and consistent with the FGA specifications: mip_era, activity_id, product, institution_id, source_id, experiment_id, sub_experiment_id (set to "none" except for DCPP[1] experiments), variable_id, table_id, variant_label, frequency, realm, grid_label.

The above are likely to be used in faceted searches; some are used in generating file and directory names.

3. Provenance etc.

The following global attributes must satisfy the objective test listed (further details in the specification may also need checking, but will not be requirements for publication):

Test	**Global Attribute tested**	**Objective test**
Persistent Identifier	tracking_id	Must be syntactically correct, i.e. of the form **hdl:21.14100/<uuid>** [2]
Link to documentation	further_info_url	Must be syntactically correct.[3]
Date of file generation	creation_date	Must be syntactically correct: "YYYY-MM-DDTHH:MM:SSZ" (e.g. "2005-01-01T18:00:00Z").

In addition, the grid_resolution should be checked for conformance with the CV[4] because it may be used in faceted searches. In addition the Conventions and data_specs_version

[1] The sub_experiment_id for DCPP runs carries the start year in the format "sYYYY" -- [still to be finalised: "sYYYYMMDD" is under consideration].

[2] The specifications state that the tracking_id should be unique. This means that it should be changed every time a new version of a file is published.

[3] http://es-doc.org/cmip6FurtherInfo/<document identifier> **[further details to be added when available]**.

[4] Subject to finalisation of the CV.

attribute should be checked because if they are incorrect, software might misinterpret all the other attributes. The global attributes parent_time_units, branch_time_in_child, and branch_time_in_parent must be present when the experiment has a defined parent. Finally, the realization_index, initialization_index, physics_index, and forcing_index should be checked for consistency with the variant_label.

4. Variable name, dimensions and attributes.

- The variable name must be the same as variable_id and the first segment of the file name, and must be a valid CMIP6 variable name listed in the data request. It must be found in the table recorded in table_id.
- The variable's dimensions must be as specified in the data request (in some cases there may be more than one table record matching the variable name: the file must match one of the specifications), matching the name and order (except that, (1) in some cases there are permitted variants on the names of spatial dimensions .. see section 6 below);
- Variable Attributes: the following variable attributes must match the data request specifications: units, standard_name, coordinates, missing_value, _FillValue;
- Coordinate variable attributes, : units, standard_name, axis, positive (only when axis="Z").
- Variable and coordinates must be of the specified type.

5. Temporal Dimension

- The time access has name "time" and units of "days since ….."[5];
- The time variable will carry a valid CF calendar attribute;
- The time axis values within a file should increase monotonically in steps which are consistent with the declared data frequency.[6][7]

6. Irregular and unstructured grids

- Data on unstructured grids must declare the UGRID convention in the "Conventions" global attribute. There are no automated checks for this convention at present, and this form of data is new to CMIP.
- Data on irregular grids must specify the grid coordinates in compliance with the CF convention following one of the following models:
    - Rotated lat-lon grid: dimensions must be rlat, rlon with standard_name, units and long_name as in data request;
    - Projection: dimension must be x,y with standard_name, units and long_name as in data request [see comment above];

[5] The specification also requires that the time units be constant across all data associated with a particular experiment. This is not enforced here because it may be difficult to verify in distributed processing chains, but data providers should check if possible.

[6] Subject to checks on the time taken to read time axis from NetCDF files.

[7] Time steps should be constant, except for monthly data which may have steps varying from 28 to 31.

- Generic indexed: dimensions must be i,j,k,l,m units and long_name as in data request [see comment above].

7. CF Compliance

Data files should be CF compliant, except where there are known exceptions for specific variables.[8]

Rationale and discussion

Having robust publication requirements will avoid some of the most serious problems encountered in CMIP5, such as data with invalid file names, while ensuring that publication can work smoothly;

Some deliberate omissions (which others might want to include):

- Checks on variable ranges: we have sign errors and errors in units (e.g. percentages instead of fractions or vice versa). While these are a significant problem for users they do not impede the data distribution service. Reliable detection for all variables is difficult.
- Checks on consistency of time units between files: when all the files are in one place it is easy to check that the units attribute of the time dimension is the same across all files in an experiment, but the data publication workflows will generally be more complex than “assemble files from a given experiment on a server → check files → publish files”, so we cannot assume that it is easy to check consistency with other data from the same experiment. It would be possible to check if the data request requires a specific value for each experiment …
- Checks on bounds attributes and associated variables: such checks are valuable, but non-compliance does have to impede data publication;

Software for data verification

Some checks will be carried out within CMOR and the ESGF publication change. The following tools provide independent checks (the last two cannot be configured for CMIP6 until these requirements are agreed).

CF Convention checker

For checking compliance with the CF Convention

[8] A small number of CMIP5 variables had specifications in the data request which were incompatible with the CF convention. These specific problems will be corrected, but it is possible that new problems of the same kind will be introduced among the range of complex new variables in the CMIP6 data request.

DKRZ Quality Control

An extensive suite of tests, tests all requirements listed in this document (not yet .. subject to finalisation of requirements and configuration), plus many additional aspects of the specifications and general data quality issues.

CEDA Compliance Checker

A more limited set of tests, with emphasis on portability and user friendly reporting. Tests all requirements listed in this document (not yet .. subject to finalisation of requirements and configuration)

CMOR Checker

Checks metadata and data against all requirements listed in this document, and more. This checker can identify errors before the files are written by CMOR, or it can be used independent of CMOR to perform checks on files already written (or as part of the ESGF publication procedure).