CMIP5 and AR5 Data Reference Syntax (DRS)

Karl E. Taylor, V.Balaji, Steve Hankin, Martin Juckes, Bryan Lawrence

Version 0.19: 29 June, 2009.

1 Introduction

1.1 Scope

This document provides a common naming system to be used in files, directories, metadata, and URLs to identify datasets wherever they might be located within the distributed CMIP5 federation.

1.2 Context:

The CMIP5 archive will be distributed between several centers using different storage architectures. As far as possible these differences should be hidden from the user.

The data reference syntax (DRS) should be sufficiently flexible to cover all the services which the archive might wish to offer, even though resource limitations may restrict the services which are actually delivered within the CMIP5 time frame. The DRS needs to take account of the user resources (usually a file system based data store) and the software to be used by the archive (such as OPeNDAP). The context in which the system will be used will require a compromise between brevity and clarity but there should be no ambiguity and easily accessible expansions of all terms.

1.3 Purpose

The Data Reference Syntax (DRS) should provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive and of files delivered to users. The DRS should make use of controlled vocabularies to facilitate documentation and discovery. Providing users with data in files with well structured names will facilitate management of the data on the users' file systems and simplify communication among users and between users and user support. The controlled vocabularies will be useful in developing category-based data discovery services. The elements of the controlled vocabularies will occur frequently in software and web pages, so they should be chosen to be reasonably brief, reasonably intelligible, and avoid characters which may cause problems in some circumstances (e.g. "/", "(", ")").

1.4 Use Case and Requirements

There are 5 specific use cases which the DRS must support:

1. Those responsible for replicating data within the CMIP5 federation should be able to exploit the DRS to guide what needs to be replicated, and to where.
2. Those responsible for the federation catalogues should be able to use the DRS to identify to catalogue users unambiguously which replicants are available for download or for on-line access (such as OPeNDAP).

3. Those responsible for the archives should be able to use the DRS to advise on a file layout (if they use a file systems as their storage management system),
4. Users should be able to use modify download scripts in a completely transparent manner, so that for example, a slow wget from one site, can be repeated (or finished) using a script in which only the hostname part of the DRS has been changed.
5. The names of the core datasets should be predictable enough that, for example, a user having found and downloaded or accessed data on-line from one model simulation can modify their scripts to download or access another model and/or simulation with only knowledge of the relevant controlled vocabulary terms (model and/or simulation names).

In addition:

6. The DRS should be sufficiently extensible to describe variables and time periods beyond those defined in the CMIP5 core.

2 Definitions

2.1 Atomic dataset:

Model archives consist of collections of "atomic datasets", defined as follows:

The collection of data that is output from a single model run and characterized by sharing a single activity, institute, model, experiment/scenario, data frequency, modeling-realm, variable name, local ensemble member, and version.

The definition is intended to provide a well founded naming system to record archive contents in a structured way. An atomic dataset consists of one variable (field). For each variable the atomic dataset contains the entire spatio-temporal domain, with one value at each included time and position. The "atomic datasets" may be very large entities, with up to 1000 years of daily model output – it is not intended that they represent the chunks of data which can usefully be put into single files. The first six components (activity, institute, model, scenario/experiment, data frequency and variable name) should all come from controlled component vocabularies.

This definition explicitly does not address the fact that model intercomparison experiments may (and in the case of CMIP5, do) require multiple temporal portions to be archived for some experiments – these are notionally subsets of the atomic dataset.

2.2 Controlled Components

PCMDI is the authoritative source for these controlled vocabularies, a draft of these appears in the appendix of this document.

Activity: This component will allow the DRS to be extended to other model intercomparisons and other data archives. For CMIP5 all the archived data will be discoverable under the

"CMIP5" activity. In some cases there may be other activities (e.g., CFMIP and PMIP), which have been coordinated with CMIP5, so these activities may be cross-referenced or aliased with CMIP5 for certain portions of the CMIP5 archive.

Institute: This identifies the institute responsible for the model results. eg. UKMO. For CMIP5 the institute name will be defined by the research group at the institute, subject to approval by PCMDI.

Model: This identifies the model used (e.g. HADCM3, HADCM3-233). The modeling group will assign this name, which would be expected to include a version number (usually truncated to the nearest integer).

Experiment: This identifies either the experiment or both the experiment family and a specific type within that experiment family. In CMIP5, for example, "rcp45" refers to a particular experiment in which a "representative concentration pathway" (RCP) has been prescribed which leads to an approximate radiative forcing of 4.5 W m^{-2}. As another example, "historicalGhg" is a "historical" run with "Ghg" forcing. In this latter case, "historical" is the experiment family and "ghg" is the specific type of historical run. These experiment names are not freely chosen, but come from controlled vocabularies (see the tables at the end of this document).

Frequency: This indicates the interval between individual time-samples in the atomic dataset. For CMIP5, the following are the only options: "yr", "mon", "day", "6hr", "3hr", "30min", "monClim" (cimatological monthly mean) or "fx" (fixed, i.e., time-independent).

Modeling-realm: This indicates the high level modeling component which is particularly relevant. For CMIP5, choose from: "atmos", "ocean", "land", "landIce", "seaIce", "aerosol" "atmosChem", ocnBgchem (ocean biogeochemical). Note that sometimes a variable will be equally (or almost equally relevant) to two or more "realms", in which case the atomic dataset might be assigned to a primary "realm", but cross-referenced or aliased to the other relevant "realms".

Variable name: For CMIP5, each variable is uniquely identified by a combination of two strings: 1) a name associated generically with the variable (typically, as recorded in the netCDF file – e.g., tas, pr, ua), and 2) the name of the CMOR variable table (e.g., Amon, da, aero) in which the variable appears. Together these two strings ensure that the variable is uniquely defined. There are some applications in which the second string might be unnecessary (e.g., for CMIP5, this will be omitted from the directory structure). Note that for CMIP5 a hyphen ('-') is forbidden to be included anywhere in the first string.

Ensemble member (r<N>i<M>p<L>)

This distinguishes among closely-related simulations by a single model. Different simulations that are equally likely outcomes for a particular simulation (i.e., they typically differ only by being started from equally realistic initial conditions) are distinguished by different integer values of "N". CMIP5 historical runs initialized from different times of a control run, for example, would be identified by "r1", "r2", "r3", etc.). The data supplier must assign a realization number to each atomic dataset..

Models used for forecasts that depend on the initial conditions might be initialized from observations using different methods. Simulations resulting from initializing a model with different *methods* should be distinguished by assigning different values of "M" in the "initialization method indicator (i<M>). For CMIP5 this indicator might in some cases be needed to distinguish among runs performed as part of the suite of decadal prediction experiments (1.1-1.6). For these experiments, the data supplier must assign an initialization method number to each atomic dataset; for all other experiments this number is irrelevant and should be omitted from the description. A key that defines the various initialization methods should be made available, so that a user can learn which initialization method is associated with each value of M.

If there are many, very closely related model versions, generally referred to as a perturbed physics ensemble (e.g., QUMP or climateprediction.net ensembles), then these should be distinguishable by a "perturbed physics" number, p<L>, where the integer is uniquely associated with a particular set of model parameters (e.g., r3p78 is a third realization of the seventy-eighth version of the perturbed physics model). A key that defines the various model versions (i.e., a table) should be made available, so that a user can learn which set of parameter values is associated with each value of L.

Version number (vN)

The version number will be 'v' followed by an integer, which uniquely identifies a particular version of the output (e.g., perhaps distinguishing between an original version of the output that might have been found to be flawed in some respect--perhaps due to some improper post-processing procedure-- and a subsequent version in which the data were corrected).

2.4 Extended Path

Note that thus far we have not yet considered datasets which might be spatio-temporal subsets. We expect these to exist both as files in the archive as well as virtual files (that is, URLs representing aggregated time series of files that are accessible by services such as OPeNDAP). The DRS supports the organization of data using such subsets, owever, these indicate "parts" of an atomic dataset, hence they were not included in the definition of atomic dataset above.

Temporal Subsets: Time instants and periods (N1(-N2))

Time instants will be represented by 'yyyy[mm[dd[hh]]][-clim]', where 'yyyy', 'mm', 'dd', 'hh' are integer year, month, day and hour respectively, and enough (and just enough) of the suffixes should be added to unambiguously resolve the interval between time-samples contained in the file or virtual file URL. (For example, monthly mean data would include "mm", but not "dd" or "hh"; daily data would include "mmdd", but not "hh".) The optional "-clim" is appended when the file contains a climatology (e.g., a file with sampling frequency of "mo" with the time designation 196001-198912-clim) represents the monthly mean climatology (12 time values) computed for the period extending from 1/1960-12/1989). Note that the DRS does not explicitly indicate the calendar, but the calendar will be indicated by one of the attribute in each netCDF file.

Geographic Subsets

It is not (currently) likely that geographical subsets described by bounding boxes will be stored in the archive, however, subsets by named location may be. Where these appear in the extended

Path, they should appear last as gXXXXX where XXXXX is a name from a specific gazetteer (which is yet to be selected).

2.5 Permitted Characters.

The character set permitted in the components needs to be restricted in order that strings formed by concatenating components can be parsed. For the purposes of this scoping exercise, it will be assumed that the components will be used in URLs, punctuated by "/", "=", ":", and "?", and in the names of files delivered to users, punctuated by "." and "_". Thus, none of these characters can be permitted within the component values and so the permitted characters will be: a-z, A-Z, 0-9, and "-".

3. Using the DRS Syntax

Here are three use cases for the DRS syntax: in URLs, for a directory layout, and in filenames.

3.1 URL syntax.

When the DRS is used in a URL, we would expect the URL to comprise a hostname, the atomic dataset name, possibly an extended path name, and possibly a service endpoint name. That is, we would expect to see usage like:

http://<hostname>/<activity>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable>/<ensemble member>/<version>/[<endpoint>],

where square brackets enclose optional elements (in this case, only the service endpoint).

Where no service endpoint appears, it should be expected that an HTTP GET on the URL will return the NetCDF data. (Currently there is no CMIP5 controlled vocabulary for endpoints, when one appears it will have values which encompass services such as OPeNDAP and WCS etc.)

Note that ensemble member and version number are mandatory, to ensure that if subsequent versions or ensemble members appear, there is no possibility of ambiguity as to what data is referenced at a given URL.

Should replace the following with "real" examples

http://badc.nerc.ac.uk/activity/institute/model/experiment/frequency/realm/varname/r1/v1/

or

http://badc.nerc.ac.uk/activity/institute/model/experiment/frequency/realm/varname/r1/v1/extended_path/

or

http://badc.nerc.ac.uk/activity/institute/model/experiment/frequency/realm/varname/r1/v1/extended_path/service_endpoint

or

http://badc.nerc.ac.uk/activity/institute/model/experiment/frequency/realm/varname/r1/v1/service_endpoint

Controlling the vocabulary for service endpoints is beyond the scope of this document, but will be a necessary part of the distributed URL design, and impact on what appears in catalogues.

However, we might expect that without a service endpoint, dereferencing these URLs will return either netcdf data, or catalogue entries. (Examples of service endpoints, might be : las, opendap, wcs, wms, wfs etc).

(Note that "hostnames" will probably be intuitional virtual hostnames, rather than individual system names, but either way, will need to be present in catalogues).

BNL Note: actually, once one starts considering service endpoints there is a strong argument that the variable name should be after the realization and version numbers, allowing one to construct service endpoints which serve multiple variables

3.2 Directory Layout

For CMIP5, the directory layout could consist of the atomic dataset name laid out as in the URL syntax, followed by subdirectories – if desired – to match the extended path:

<activity>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable name>/<ensemble member>/

For example

/CMIP5/UKMO/HADCM3/decadal1990/day/atmos/tas/r3i2/

/CMIP5/UKMO/HADCM3/rcp45/mon/ocean/uo/r1/

Note that for CMIP5 the second of the two strings that define the variable name (as discussed in the section on "Controlled Components") is dropped in the directory structure. The table name ("Amon in this case) is dropped, since the "activity", the "frequency" and the "modeling realm", which already appear in the directory path, together unambiguously imply a certain table.

3.3 Filenames

Because users will download data into a file system that will usually differ from the archival directory structure (and because in some cases it aids in archive management), the filename structure should include some DRS content. For CMIP5 the filename will be constructed as follows:

filename =

*<1st part of varname>*_*<2nd part of varname>*_*<model>*_*<experiment>*_<ensemble member>_*<temporal subset>*.nc

where:

- *<1st part of varname>,* <2nd part of varname>, *<model>*, *<experiment>*, and <ensemble member> are from the atomic dataset definition,

- The < *time period*> is from the extended path definition.

Example:

tas_Amon_HADCM3_ historical_r1_185001-200512.nc