

CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies

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

Version 0.26

24 March 2010.

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. It defines controlled vocabularies for many of the components comprising the data reference syntax (DRS).

1.2 Context:

The CMIP5 archive will be distributed among 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. “/”, “(”, “)”).

Formatted: Font: 18 pt

Formatted: Centered

Formatted: Centered, Level 1, Space Before: 0 pt, After: 0 pt

Deleted: CMIP5 and AR5 Data Reference Syntax (DRS)

Formatted: Font: 18 pt

Formatted: Font: 18 pt, No underline

Formatted: Centered

Formatted: Font: 14 pt

Deleted: 5

Deleted: :

Deleted: 28

Deleted: February

1.4 Use Case and Requirements

There are 6 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 define a logically structured file layout (if they use file systems as their storage management system).
4. Users should be able to 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 using a script can modify that script to download or access another model and/or simulation with only knowledge of the relevant controlled vocabulary terms (in this case, the model and/or simulation names).
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 constituting a single *product* saved from a single run which is uniquely characterized by a single *activity*, *institute*, *model*, *experiment*, *data sampling frequency*, *modeling-realm*, *variable identifier*, *ensemble member*, and *version number*.

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 1000 years of daily model output or more – it is not intended that they necessarily represent the chunks of data which can practically be put into single files. The first eight components (*product*, *activity*, *institute*, *model*, *experiment*, *frequency*, *modeling-realm*, and *variable identifier*) should all come from controlled component vocabularies, and the structure for the last two components is also controlled.

Formatted: Keep with next

This definition explicitly does not distinguish between the different temporal portions of an experiment – these are notionally subsets of the atomic dataset – even though in CMIP5 different portions of the same experiment may be assigned different priorities (e.g., extensions of some of the future scenario runs to the 22nd and 23rd centuries).

2.2 Component Definitions and Controlled Vocabularies

After seeking community input, PCMDI has final authority for defining the controlled vocabularies that together with the component categories comprise the DRS. These components and vocabularies are defined below. (See also Appendix 1.1 and Appendix 1.2.).

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.

Product: This allows distinctions between various types of model data products. For CMIP5 the only initially permissible options for product are “output”, which refers to all the model output published, and “requested”, which refers to the subset of model output specified by the CMIP5 data request¹. The “requested” product type allows users (and data managers) to focus on the subset of the complete output that is likely to be available from most of the models. Note that an atomic dataset defined under the “requested” classification will also be part of (or in some cases identical to) an atomic dataset defined under the “output” classification. There may also be “output” atomic datasets that do not include any “requested” data. [WILL POSSIBLY MODIFY THE ABOVE IF WE DON’T NEED TO KNOW ABOUT “REQUESTED”].

It is likely that various products derived from this output will be produced subsequently which could be identified by a different term (e.g., “derived” or “processed”), but this is not part of the current DRS. Alternatively, the DRS might be modified in the future to include an additional component (perhaps called “Processing”), which could be used to distinguish between “raw” (i.e., the model output itself) and “derived” products of various kinds (e.g., climatologies, zonal means, EOF analyses, etc.).

Institute: This identifies the institute responsible for the model results (e.g. UKMO), and it should be as short as possible. For CMIP5 the institute name will be suggested by the research group at the institute, subject to final authorization by PCMDI .

Model: This identifies the model used (e.g. HADCM3, HADCM3-233). Subject to certain constraints imposed by PCMDI, the modeling group will assign this name, which might 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

¹ See [CMIP5 list of requested model output](http://cmip-pcmdi.llnl.gov/cmip5/data_description.html): http://cmip-pcmdi.llnl.gov/cmip5/data_description.html

Deleted: -----Page Break-----

Deleted: standard

Deleted: description from PCMDI

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 “greenhouse gas” forcing. In this latter case, “historical” is the experiment *family* and “GHG” is used to designate the specific *type* of historical run. These experiment names are not freely chosen, but come from controlled vocabularies defined in the Appendix 1.1 of this document under the column labeled “Short Name of Experiment”.

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”, “subhr” (sampling frequency less than an hour), “monClim” (climatological monthly mean) or “fx” (fixed, i.e., time-independent).

Modeling-realm: This indicates the high level modeling component which is particularly relevant. For CMIP5, permitted values are: “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 identifier: For CMIP5, each variable is uniquely identified by a combination of two strings: 1) a “variable name” associated generically with the variable (typically, as recorded in the netCDF file – e.g., tas, pr, ua), and 2) a “variable table” name (e.g., Amon, day, aero) in which the variable appears. These two components of the variable name are defined in the so-called “standard_output” spreadsheet found at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that for CMIP5 a variable is also uniquely defined by the DRS component “frequency” and the variable name alone (without reference to a specific table). Note that within CMIP5 variable names, hyphens (‘-’) are forbidden.

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

This triad of integers (N, M, L), formatted as shown above (e.g., “r3i1p21”) distinguishes among closely-related simulations by a single model. All three are required even if only a single simulation is performed.

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 positive 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. It is recommended that the numbers be assigned sequentially starting with 1. In CMIP5 time-independent variables (i.e., those with frequency=“fx”) are not expected to differ across ensemble members, so for these N should be always assigned the value zero (“r0”).

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 positive integer 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). The data supplier must assign an initialization method number to each

atomic dataset. It is recommended that the numbers be assigned sequentially starting with 1. In CMIP5 time-independent variables (i.e., those with frequency="fx") are not expected to differ across ensemble members, so for these M should always be assigned the value zero ("i0"). A key (i.e., a table) 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 positive integer value of L is uniquely associated with a particular set of model parameters (e.g., r3i1p78 is a third realization of the seventy-eighth version of the perturbed physics model). The data supplier must assign a physics version number to each atomic dataset. It is recommended that the numbers be assigned sequentially starting with 1. In CMIP5 time-independent variables (i.e., those with frequency="fx") are not expected to differ across ensemble members, so for these L should always be assigned the value zero ("p0"). A key (i.e., a table) that defines the various model versions should be made available so that a user can learn which set of parameter values is associated with each value of L.

Note that for a single model and experiment (i.e., across all members of the ensemble), the values assigned to N, M, and L (which together define an individual ensemble member) should each be uniquely associated with, respectively, a specific initial condition, initialization method, and perturbed physics version. Thus, these numbers will be consistent across all members of an ensemble. For example the two different ensemble members, r3i1p7 and r3i1p8, should both be initialized from *exactly the same initial conditions using the same method* (because the "r and values" are identical) although the subsequent evolution of the simulations will presumably differ since they were produced by two different perturbed versions of the same model. There may be cases where "gaps" could be found in the list of ensemble members. If, for example, two different initialization procedures were used, but the second procedure was tested with only a subset of the initial condition cases of the first procedure (say, every other case). Then the list of ensemble members would look like: r1i1p1, r2i1p1, r3i1p1, r4i1p1, r5i1p1, r6i1p1, ..., r1i1p2, r3i1p2, r5i1p2, ...

Another requirement for CMIP5 is that each so-called RCP (future scenario) simulation should be assigned the same realization integer as the historical run from which it was initiated. This will allow users to easily splice together the appropriate historical and future runs. [A similar convention should be followed, when appropriate, with other simulations \(e.g., the decadal simulations\).](#)

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 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 specification of such subsets, however, these represent only “parts” of an atomic dataset, and hence they were not included in the definition of atomic dataset above.

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

Time instants and periods will be represented by ‘yyyy[mm[dd[hh][mm]]][-clim]’, where ‘yyyy’, ‘mm’, ‘dd’, ‘hh’, ‘mm’ are integer year, month, day, hour, and minute, 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”, “hh”, or “mm”; “subhr” data would include all suffixes.) The optional “-clim” is appended when the file contains a climatology. For example, a file with sampling frequency of “mo” and 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 specify the calendar type (e.g., Julian, Gregorian), but the calendar will be indicated by one of the attributes in each netCDF file.

Geographic Subsets

It is (currently) unlikely that geographical subsets described by bounding boxes will be stored in the archive, but subsets by named location might 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. Other characters will also be excluded at this time, 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>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable identifier>/<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.)

Formatted: Keep with next

Deleted: and

Deleted: or

Deleted: daily

Deleted: “mmdd”, but not “hh”

Note that ensemble member and version numbers 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/product/institute/model/experiment/frequency/realm/
varname/r1/v1/`

or

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

or

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

or

`http://badc.nerc.ac.uk/activity/product/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 identifier should be after the realization and version numbers, allowing one to construct service endpoints which serve multiple variables

3.2 Directory Layout

For CMIP5, certain software will assume a directory layout as follows:

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

For example

`/CMIP5/output/UKMO/HADCM3/decadal1990/day/atmos/tas/r3i2p1/v1/`

`/CMIP5/output/UKMO/HADCM3/rcp45/mon/ocean/uo/r1i1p1/v3/`

Links to the latest version of all files constituting the dataset will appear directly in the <ensemble member> directory. The files themselves will be found a level below this in a particular <version number> subdirectory. Note that for CMIP5 the second of the two strings that identify the variable (i.e., the “variable table” discussed in the section on “Controlled Components”) does not appear in the directory structure, since the “activity”, the “frequency” and the “modeling realm”, which already do appear in the directory path, together unambiguously imply a certain table.

Deleted: is dropped

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 =

```
<variable name>_<variable table>_<model>_<experiment>_<ensemble member>[_<temporal subset>].nc
```

where:

- <variable name>, <variable table>, <model>, <experiment>, and <ensemble member> are from the atomic dataset definition,
- The <temporal subset> is omitted for variables that are time-independent.

Note that together, <variable name> and <variable table> constitute the “variable identifier”, which uniquely defines the variable.

Formatted: No bullets or numbering

Formatted: Font: Not Italic

Deleted: .

Example:

```
tas_Amon_HADCM3_historical_r1i1p1_185001-200512.nc
```

Appendix: Controlled Vocabularies

Appendix 1.1 Experiment Controlled Vocabulary

Coupled Model “Decadal” Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
1.1, 1.2 & 1.5	decadalXXXX*	10- or 30-year run initialized in year XXXX*	decadal hindcasts/predictions, some extended to 30 years	10-30	$\geq 3^* \geq 10$
1.3	noVolcXXXX*	volcano-free hindcasts	hindcasts but without volcanoes	10-30	≥ 3
1.4	volcIn2010	prediction with 2010 volcano	Pinatubo-like eruption imposed in year 2010	10-30	≥ 3
1.6	**	chemistry-focused runs	near-term runs with enhanced chemistry/aerosol models	10-30	1

* Replace 'XXXX' with the year in which the decadal prediction was initiated (typically near the end of year XXXX). If a run is initiated on January 1, then XXXX should be the year immediately preceding the date of initialization.

** These runs will be placed in the appropriate directory (defined by the experiment; 1.1-1.5); Experiment 1.6 differs from the others only because a different model is used, which will be indicated by a unique model name, so there is no need for a new directory for these runs.

Coupled Model Long-Term Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
3.1	piControl	pre-industrial control	coupled atmosphere/ocean pre-industrial control run	≥500	1
3.2	historical	historical	simulation of recent past (1850-2005)	156	≥1
3.4	midHolocene	mid-Holocene	consistent with PMIP, impose Mid-Holocene conditions	100	1
3.5	lgm	last glacial maximum	consistent with PMIP, impose last glacial maximum conditions	100	1
3.6	past1000	last millennium	consistent with PMIP, impose forcing for 850-1850	1000	1
4.1	rcp45	RCP4.5	future projection (2006-2300) forced by RCP4.5	95-295	1
4.2	rcp85	RCP8.5	future projection (2006-2300) forced by RCP8.5	95-295	1
4.3	rcp26	RCP2.6	future projection (2006-2300) forced by RCP2.6	95-295	1
4.4	rcp60	RCP6	future projection (2006-2100) forced by RCP6	95	1
5.1	esmControl	ESM pre-industrial control	as in experiment 3.1, but emissions-forced (with atmospheric CO2 determined by model)	250	1
5.2	esmHistorical	ESM historical	as in experiment 3.2, but emissions-forced (with atmospheric CO2 determined by model)	156	1
5.3	esmrcp85	ESM RCP8.5	as in experiment 4.2, but emissions-forced (with atmospheric CO2 determined by model)	95	1

5.4-1	esmFixClim1	ESM fixed climate 1	radiation code "sees" control CO2, but carbon cycle sees 1%/yr rise	140	1
5.4-2	esmFixClim2	ESM fixed climate 2	radiation code "sees" control CO2, but carbon cycle sees historical followed by RCP4.5 rise in CO2	251	1
5.5-1	esmFdbk1	ESM feedback 1	carbon cycle "sees" control CO2, but radiation sees 1%/yr rise	140	1
5.5-2	esmFdbk2	ESM feedback 2	carbon cycle "sees" control CO2, but radiation sees historical followed by RCP4.5 rise in CO2	251	1
6.1	1pctCO2	1 percent per year CO2	imposed 1%/yr increase in CO2 to quadrupling	140	1
6.3	abrupt4xCO2	abrupt 4XCO2	impose an instantaneous quadrupling of CO2, then hold fixed	150	≥1
7.1	historicalNat	natural-only	historical simulation but with natural forcing only	156	≥1
7.2	historicalGHG	GHG-only	historical simulation but with greenhouse gas forcing only	156	≥1
7.3	historical?***	other-only	historical simulation but with other individual forcing agents	156	≥1

*** The “?***” should be replaced with a two- or three-letter character string, which will uniquely identify the individual forcing agent that is active. Choose strings from the abbreviations given in Appendix 1.2.

Atmosphere-Only Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
3.3	amip	AMIP	AMIP (1979- at least 2008)	≥ 30	≥ 1
2.1	sst2030	2030 time-slice	conditions for 2026-2035 imposed	10	≥ 1
6.2a	sstClim	control SST climatology	control run climatological SSTs & sea ice imposed	30	1
6.2b	sstClim4xCO2	CO2 forcing	as in experiment 6.2a, but with 4XCO2 imposed	30	1
6.4a	sstClimAerosol	all aerosol forcing	as in experiment 6.2a, but with aerosols from year 2000 of experiment 3.2	30	1
6.4b	sstClimSulfate	sulfate aerosol forcing	as in experiment 6.2a, but with sulfate aerosols from year 2000 of experiment 3.2	30	1
6.5	amip4xCO2	4xCO2 AMIP	AMIP (1979-2008) conditions (experiment 3.3) but with 4xCO2	30	1
6.6	amipFuture	AMIP plus patterned anomaly	consistent with CFMIP, patterned SST anomalies added to AMIP conditions (experiment 3.3)	30	1
6.7a	aquaControl	aqua planet control	consistent with CFMIP, zonally uniform SSTs for ocean-covered earth	5	1
6.7b	aqua4xCO2	4xCO2 aqua planet	as in experiment 6.7a, but with 4XCO2	5	1
6.7c	aqua4K	aqua planet plus 4K anomaly	as in experiment 6.7a, but with a uniform 4K increase in SST	5	1
6.8	amip4K	AMIP plus 4K anomaly	as in experiment 3.3, but with a uniform 4K increase in SST	30	1

Appendix 1.2 Controlled Vocabulary for Abbreviated “Forcing” Descriptors

Abbrev.	Forcing Description	Abbrev.	Forcing Description
Nat	natural forcing (a combination, not explicitly defined here, that might include, for example, solar and volcanic)	LU	land-use change
Ant	anthropogenic forcing (a mixture, not explicitly defined here, that might include, for example, well-mixed greenhouse gases, aerosols, ozone, and land-use changes).	SI	solar irradiance (note: SI is “S” followed by a lower case “L”, not an upper case “I”)
GHG	well-mixed greenhouse gases (a mixture, not explicitly defined here)	VI	volcanic aerosol (note: VI is “V” followed by a lower case “L”, not an upper case “I”)
SD	anthropogenic sulfate aerosol, accounting only for direct effects	SS	sea salt
SI	anthropogenic sulfate aerosol, accounting only for indirect effects	Ds	Dust
SA (= SD + SI)	anthropogenic sulfate aerosol direct and indirect effects	BC	black carbon
TO	tropospheric ozone	MD	mineral dust
SO	stratospheric ozone	OC	organic carbon
Oz (= TO + SO)	ozone (= tropospheric and stratospheric ozone)	AA	anthropogenic aerosols (a mixture of aerosols, not explicitly defined here)