CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies

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

Version 1.3.1

6 June 2012

Karl Taylor 1/15/12 10:02 AM
Deleted: 2.2

Karl Taylor 1/15/12 10:02 AM
Deleted: 21 November 2011

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 archive 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 that the archive might wish to offer, even though resource limitations may restrict the services that 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 6 specific use cases which the DRS must support:

1. Those responsible for replicating data within the CMIP5 archive 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:

> **Atomic dataset definition: a subset of the output saved from a single model run which is uniquely characterized by a single *activity, product, institute, model, experiment*, data sampling *frequency*, *modeling realm*, *variable name, MIP table, 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 values reported at each included time and location. An "atomic dataset" may be a very large entity, with 1000 years of daily model output or more; it does not necessarily represent a chunk of data that can practically be put into a single file. The first nine components (*activity, product, institute, model, experiment, frequency,*

modeling realm, *variable name* and *MIP table*) should all come from controlled component vocabularies, and the structure for the last two components is also controlled.

2.2 Publication-level dataset

When applied to the CMIP5 experiment definition the atomic dataset definition above leads to millions of atomic datasets. This level of granularity is too fine for the data management technologies employed for CMIP5, therefore atomic datasets are aggregated into "publication-level" datasets[1] containing all variables for a single combination of other DRS components.

> **Publication-level dataset definition: The collection of atomic datasets which share a single combination of all DRS component values except *variable name* but which might include only selected time intervals (i.e., not necessarily the entire temporal domain) of the contributing atomic datasets. The publication-level dataset therefore represents, in general, an intersection of several atomic datasets.**

Note that the *version number* component is effectively a property of publication-level datasets.

2.3 Component Definitions and Controlled Vocabularies

Karl Taylor 1/15/12 10:03 AM
Deleted: 2

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 identifies the model intercomparison activity or other data collection activity. For CMIP5 all the archived data will be discoverable under the "CMIP5" activity. For "Transpose AMIP", the data will be archived under the "TAMIP" 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 currently has four options: "output", "output1", "output2", and "unsolicited". For CMIP5, files will initially be designated as "output" or "unsolicited". Subsequently, data from the requested variable list will be assigned a version (see below) and placed in either "output1" or "output2". Variables not specifically requested by CMIP5 will remain designated "unsolicited". In some cases a continuous sequence of model data will be split between "output1" and "output2" in order to facilitate archive management. Note that although output of some variables is requested only for limited time-periods, if output of those variables is made available for other time periods, it will also be treated as "output", not as "unsolicited".

Karl Taylor 6/6/12 9:30 AM
Deleted:

---

[1] Publication-level datasets have previously been referred to as "Realm-level datasets" in internet communications related to CMIP5 such as email lists and wiki pages.

[2] For CMIP5 the precision of the time values corresponding to each MIP table is as follows: "yyyy" for table Oyr;

It is likely that various data 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.

Institute 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. This name may differ somewhat from the official CMIP5 institute_id (recorded as a global attribute in CMIP5 output files), which should be used to identify models in journal articles. [The official institute_id might, for example, include characters such as a blank, a period, or a parenthesis, which are not allowed in the DRS "institute" component.]

Model 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). This name may differ somewhat from the official CMIP5 model_id (recorded as a global attribute in CMIP5 output files),, which should be used to identify models in journal articles. [The official model_id might, for example, include characters such as a blank, a period, or a parenthesis, which are not allowed in the DRS "model" component.] The model identifier will normally change if any aspect of the model is modified (e.g., if the resolution is changed). An exception may be made if the modifications to the model are clearly implied by the experiment design. If, for example, a coupled atmosphere-ocean model performs an AMIP simulation (which clearly implies prescribed SSTs and sea ice, rather than a fully interactive ocean), then the name may not necessarily be modified. Another exception is when closely-related "perturbed physics" versions of a model are run, in which case the different model versions can be uniquely identified by assigning each a different "p" value in defining the "ensemble member" (described below).

Experiment 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 specified which leads to an approximate radiative forcing of 4.5 W m^{-2}. As another example, "historicalGHG" is a simulation of the historical" period, but with forcing other than anthropogenic "greenhouse gas" forcing suppressed. 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". Note that in some cases there will be slight variations of the same experiment (e.g., different simulations performed within the historicalMisc family might be forced with different individual forcings or suites of forcings, as discussed further under "Ensemble member" below).

Frequency 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). These are specified for each variable in the "standard_output" spreadsheet found at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that for CMIP5, quantities derived from an atomic dataset of a given frequency will be assigned the same frequency, even

in the case when a time-average has been performed. (See example under section 2.4 involving time averages.)

Modeling realm indicates which high level modeling component is of particular relevance for the dataset. For CMIP5, permitted values are: "atmos", "ocean", "land", "landIce", "seaIce", "aerosol" "atmosChem", ocnBgchem (ocean biogeochemical). These are specified for each variable in the "standard_output" spreadsheet which can be accessed at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. 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 and the MIP table component of the DRS (defined next) identify the physical quantity and often imply something about the sampling frequency and modeling realm. For CMIP5 the variable name and MIP table for requested output appear in the "standard_output" spreadsheet available at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Monthly mean surface air temperature, for example, has a "variable name" of "tas" and is found in the "Amon" MIP table., Note that hyphens (-) are forbidden in CMIP5 variable names.

MIP table: See description under the "variable name" component directly above. For CMIP5 each MIP table contains fields sampled only at a single frequency (although in the case of monthly mean data the DRS will place some of the monthly means in the "mon" DRS frequency category and others in the monClim DRS frequency category, as appropriate).

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.

The so-called "realization" number (a positive integer value of "N") is used to distinguish among members of an ensemble typically generated by initializing a set of runs with different, but equally realistic, initial conditions. 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 generally recommended that the numbers be assigned sequentially starting with 1 (but other recommendations, specified below, may override this recommendation). 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 invariably assigned the value zero ("r0"). For TAMIP ("the Transpose AMIP activity), the "realization" number is used to distinguish among the 16 members of each of 4 ensembles (one for each of 4 "seasons") generated from different observed conditions, spaced 30 hours apart. So, for example, the 16-member ensemble of runs initialized at 00Z on 15 Oct 2008, 06Z 16 Oct 2008, 12Z 17 Oct 2008, and so-on, would be assigned "r1", "r2", "r3", etc.

Models used for forecasts that depend on the initial conditions might be initialized from observations using different methods or different observational datasets. These 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 in the decadal-prediction suite of 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 invariably be assigned the value zero ("i0"). A key (i.e., a table) should be made available that associates each value of M with a particular initialization method and/or observational dataset.

If there are many closely related model versions, which, as a group, are 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). If there are different "forcing" combinations prescribed in experiment 7.3 in CMIP5 (the "historicalMisc" runs), then each of these different runs are also assigned different values of L (in "p<L>"). Note that 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) should be made available that associates each value of L with a particular set of model parameter values and/or, in the case of the "historicalMisc" experiment, a particular suite of "forcing" agents.

Note that for a single model and experiment N, M, and L should be interpretable independently; for all members of the ensemble, the correspondence between the values of N, M, and L and the simulation characteristics they represent should be consistent. 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 "i" values are identical) although the subsequent evolution of the simulations will presumably differ since they were produced by two different "perturbed physics" versions of the same model. Note that there may be cases where "gaps" could occur 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, …, r1i2p1, r3i2p1, r5i2p1, …

A recommendation for CMIP5 is that each so-called RCP (future scenario) simulation should when possible 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. Thus, for example, suppose a 3-member ensemble of historical runs of a model exists, and a single rcp45 simulation was produced, initialized from the third member of the historical ensemble. The recp45 simulation would be designated "r3" (rather than "r1"), even though it is the only existing ensemble member, in order to indicate that it was spawned from member 3 of the historical ensemble. 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 a publication-level dataset (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). For CMIP5 the version number is supposed to reflect the date of puclication: for example, "v20100105" for a version provided on 5th January 2010. Software

Karl Taylor 6/6/12 12:27 PM
Deleted: the output

Karl Taylor 6/6/12 1:31 PM
Deleted: T

Karl Taylor 6/6/12 1:32 PM
Deleted: will be generated from the date,

Karl Taylor 6/6/12 1:32 PM
Deleted: e.g

interpreting version numbers should not, however, assume the integer has invariably been correctly encoded (e.g., sometimes a single digit number might appear as in "v3").

Version numbers are assigned to publication-level datasets (and therefore the version generally applies to multiple atomic datasets). The version number of a publication-level dataset (and all the atomic datasets in it) is updated when:

a) any file included in the publication-level dataset is modified, replaced, or removed, or

b) an additional file is added to the publication-level dataset.

2.4 Extended Path

Note that thus far we have not considered datasets that contain spatio-temporal subsets or means. 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 or means, however, these represent or are derived from only "parts" of an atomic dataset, and hence they were not included in the definition of atomic dataset above.

Temporal subsets or means: Time instants and periods (N1-N2)

Time instants or periods will be represented by a construction of the form "N1-N2", where N1 and N2 are of the form 'yyyy[MM[dd[hh[mm[ss]]]]][-suffix]', where 'yyyy', 'MM', 'dd', 'hh' 'mm' and 'ss' are integer year, month, day, hour, minute, and second, respectively, and the precision with which time is expressed must unambiguously resolve the interval between time-samples contained in the file or virtual file. Often, these times would be expressed with as little precision as necessary to resolve the interval, but this is not a requirement. For example, monthly mean data would normally include "yyyy" and "MM", but not "dd", "hh", "mm" or "ss"; "subhr" data would include complete time precision if at least some of the time samples might not fall on minute marks (e.g., samples reported every 7.5 minutes).[2] If only a single time instant is included in the dataset, N2 may normally be omitted, but for CMIP5 N2 is required and in this case would be identical to N1.

The optional "-suffix" can be included to indicate that the netCDF file contains a climatology (suffix = "-clim") or a single time mean, for example, over multiple years (suffix = "-avg"). 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. As another example, consider a file containing a single time-average, based on daily samples for the two week period from February 1, 1971 through February 14, 1971. In this case the frequency for the dataset would be "day" (because the average is based on daily samples), and the suffix would be "19710201-19710214-avg".

[2] For CMIP5 the precision of the time values corresponding to each MIP table is as follows: "yyyy" for table Oyr; "yyyyMM" for tables Oclim, Amon, Omon, Lmon, Limon, Oimon, Aero, cfMon, and cfOff; "yyyyMMdd" for tables day and cfDay; "yyyyMMddhhmm" for tables 6hrLev, 6hrPlev, 3hr, and cf3hr; "yyyyMMddhhmmss" for table cfSites.

Karl Taylor 6/6/12 12:35 PM
Formatted: Numbered + Level: 1 + Numbering Style: a, b, c, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Karl Taylor 6/6/12 12:29 PM
Formatted: Keep with next

Karl Taylor 2/9/12 4:42 PM
Deleted: which might be

Karl Taylor 1/15/12 10:37 AM
Deleted: S

Karl Taylor 2/9/12 3:45 PM
Deleted: (

Karl Taylor 2/9/12 3:45 PM
Deleted:)

Karl Taylor 2/9/12 4:46 PM
Deleted: instants and

Karl Taylor 2/9/12 3:51 PM
Deleted: mm

Karl Taylor 2/9/12 4:44 PM
Deleted:]

Karl Taylor 1/16/12 4:43 PM
Deleted: clim

Karl Taylor 2/9/12 3:51 PM
Deleted: mm

Karl Taylor 2/9/12 3:51 PM
Deleted: and

Karl Taylor 2/9/12 3:49 PM
Deleted: enough (and just enough) of the suffixes should be added to

Karl Taylor 2/9/12 3:49 PM
Deleted: URL

Karl Taylor 2/9/12 3:56 PM
Deleted: (

Karl Taylor 2/9/12 3:54 PM
Deleted: mm

Karl Taylor 2/9/12 3:55 PM
Deleted: or

Karl Taylor 6/6/12 8:43 AM
Deleted: all suffixes

Karl Taylor 2/9/12 3:57 PM
Deleted:)

Karl Taylor 1/16/12 4:47 PM
Deleted: is appended when the file contains a climatology.

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 and spatial means

The geographical indicator is always optional, but when present it should appear last in the extended path. This indicator specifies geographical subsets described by bounding boxes (e.g. 20S to 20N and 0 to 180E) or by named regions (e.g., Pacific Ocean). The DRS specification for this indicator is a string of the form g[-XXXX][-YYYY]. The "g" indicates that some spatial selection or processing has been done (i.e., selection of a sub-global region and/or spatial averaging). The "XXXX" is optional and is either a named region (with names from a specific gazetteer, which is yet to be selected) or the bounds of a latitude-longitude rectangle (following the template defined below). The "YYYY" is optional and indicates if and what sort of spatial averaging has been performed and whether the average includes masking of certain areas within the region (e.g., masking of land areas). The "g" should be omitted unless "XXXX" and/or "YYYY" are present.

In the case of a bounding box, the bounds of the regions should be specified following the form, "latJ[pK]HJJ[pKK]HHlonM[pN]ZMM[pNN]ZZ" where the "p"s indicate a decimal point, and J, K, JJ, KK, M, N, MM, NN are integers, and the H and HH are restricted to "N" or "S" (indicating "north" or "south"), and the Z and ZZ are restricted to "E" or "W" (indicating "east" or "west"). The decimal fractions (along with their preceding p's) may be omitted if unnecessary. If all latitudes are included, then the latitude bounds can be omitted, and similarly for longitude.

The "YYYY" string is of the form "[yyy]-[zzz]" where the hyphen should be omitted unless both "yyy" and "zzz" are present. As options for "yyy", the DRS currently includes "lnd" and "ocn". The "lnd" suffix indicate that only "land" locations are considered, and the "ocn" suffix indicates that only "ocean" locations (including sea ice) are considered. As options for "zzz", the DRS currently includes "zonalavg" and "areaavg", which indicate "zonal mean" and "area mean" respectively.

Here are some examples of geographical indicators:

- "g-lat20S20Nlon170p5W130p5W" – a geographical subset defined by a bounding box (latitudes -20 to 20, and longitudes -170.5 to -130.5)
- "g-ocn-areaavg" – an average over the world's oceans.
- "g-lat20S20N-lnd-zonalavg" – a zonal average over tropical lands, covering all longitudes.

Karl Taylor 1/16/12 4:47 PM
Deleted: .

Karl Taylor 1/15/12 10:37 AM
Deleted: s

Karl Taylor 1/17/12 10:44 AM
Deleted: 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 "-". In constructing the "variable name" component of the DRS, it is recommended that the "-" be avoided since hyphens cannot be imbedded in Fortran and IDL variable names, and some users would like to maintain consistency between the DRS name and the name appearing in their code.

3. Using the DRS Syntax

The DRS component vocabularies are used in various places within the CMIP5 archive to identify digital objects. In each case there are slight variations in the encoding syntax and subset of DRS components used, reflecting the practicalities of mapping DRS concepts to different applications. Here are five examples of the use of DRS syntax: in directory structures (2 cases), in filenames, in dataset I.D.'s, and in URLs.

Karl Taylor 6/6/12 1:22 PM
Deleted: three

Karl Taylor 6/6/12 1:23 PM
Deleted: use cases for

Karl Taylor 6/6/12 1:23 PM
Deleted: the

Karl Taylor 6/6/12 1:25 PM
Deleted: , for a directory layout, and in filenames

3.1 CMOR directory structure

The standard CMIP5 output tool CMOR2[3] optionally writes output files to a directory structure mapping DRS components to directory names as:

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

For example

/CMIP5/output/MOHC/HadCM3/decadal1990/day/atmos/tas/r3i2p1/

or

/CMIP5/output/MOHC/HadCM3/rcp45/mon/ocean/uo/r1i1p1/

This structure, based on a previous version of the DRS, is incompatible with the recommended current DRS directory structure (see below). However it remains relevant as a possible structure for model output prior to transforming into the DRS directory structure.

3.2 ESGF data node directory structure

It is recommended that ESGF data nodes should layout datasets on disk mapping DRS components to directories as:

<activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<MIP table>/<ensemble member>/<version number>/<variable name>/<CMOR filename>.nc

Example:

/CMIP5/output1/UKMO/HadCM3/decadal1990/mon/atmos/Amon/r3i2p1/v20100105/tas/ tas_Amon_HADCM3_ decadal1990_r3i2p1_199001-199012.nc

[3] See the Climate Model Output Rewriter: http://www2-pcmdi.llnl.gov/cmor/documentation/

3.3 CMIP5 filename encoding

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*>_<*MIP table*>_<*model*>_<*experiment*>_
<*ensemble member*>_<*temporal subset*>[_<geographical info>].nc

where:

- <*variable name*>, <*MIP table*>, <*model*>, <*experiment*>, and <*ensemble member*> are DRS components,
- The < *temporal subset*> (along with the preceding underscore) is omitted for variables that are time-independent, and the geographical information (preceded by an underscore) is included only when needed.

Example:

tas_Amon_HADCM3_ historical_r1i1p1_185001-200512.nc

In CMIP5 there is a single exception to use of the above template. For so-called gridspec files, which describe the grids used in a model, the filename should be constructed as follows:

gridspec filename = gridspec_<*modeling realm*>_fx_<*model*>_ <*experiment*>_r0i0p0.nc

where <*modeling realm*> is now included and the variable name is replaced by "grid_spec". Note also that this is a time-independent field, so the CMIP5 table is "fx" and the ensemble member is set to "r0i0p0".

Example:

- gridspec_atmos_fx_IPSL-CM5_historical_r0i0p0.nc

3.4 Publication-level *dataset_id encoding*

Publication-level datasets are assigned an identifier *dataset_id* within THREDDS catalogs on ESGF data nodes. The CMIP5 best practices document[4] defines a publication-level *dataset_id* as:

<*activity*>.<*product*>.<*institute*>.<*model*>.<*experiment*>.<*frequency*>.<*modeling realm*>.<*MIP table*>.<*ensemble member*>

Each publication-level dataset version will have the THREDDS id:

<*activity*>.<*product*>.<*institute*>.<*model*>.<*experiment*>.<*frequency*>.
<*modeling realm*>.<*MIP table*>.<*ensemble member*>.<*version*>

[4] See CMIP5 Best Practices for Data Publication: http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/cmip5-best-practices

Karl Taylor 6/6/12 1:25 PM
Deleted: 2

Karl Taylor 2/9/12 4:52 PM
Deleted: [

Karl Taylor 1/17/12 11:10 AM
Deleted: .

Karl Taylor 6/6/12 1:24 PM
Formatted: Justified, Bulleted + Level: 1 + Aligned at: 0" + Tab after: 0.25" + Indent at: 0.25"

Karl Taylor 6/6/12 1:24 PM
Deleted:

Karl Taylor 6/6/12 1:24 PM
Deleted: *3.3 ESGF data node directory structure* ... [1]

Karl Taylor 2/9/12 4:55 PM
Deleted: P

Note that the version number assigned to the dataset by ESG is supposed to reflect the date of ESG publication, but the version will usually be assigned by the user so this cannot generally be guaranteed. The user will be instructed to provide ESG with the date that appears in the ESGF data node directory structure for the dataset being published (assuming that the directory version number is a correctly encoded date). In many cases the directory structure will be generated some days prior to publication, so the date will not in fact reflect the date of publication, but the date that the directory structure was created.

3.5 URL syntax

URLs referencing the data files will have a site dependent prefix (that may change due to site-specific data management tasks) followed by the directory structure. This directory structure should (but may not) follow the recommendations of section 3.3.