Implementation notes for the DRS

11/07/10
Martin Juckes

1 Introduction

Some aspects of the definition of DRS and it implementation with respect to replication need to be clarified. Given the specification of the replicated portion of the data in the "archive_size.xls" spreadsheet, I suggest we activate the '[WILL POSSIBLY MODIFY THE ABOVE IF WE DON'T NEED TO KNOW ABOUT "REQUESTED"]' option referred to in the paragraph defining the "product" component of the DRS in version 27.

2 Review of selection choices in *archive_size.xls*

2.1 Background

Output is requested from around 17 tables of variables – each table has a common frequency and cover a single realm [[check]]. Most of the 17 tables are further divided into sections with different spatial dimensions. The variables within each table are assigned priorities 1 (high), 2 (medium) or 3 (low). There are around 60 experiments. In some cases, output requested for the first, or first 3, ensemble member(s) is more extensive than that requested for subsequent ensemble members. In such cases, the experiment is listed twice in the "experiment" column of the archive size estimate, though it will, of course, only appear once in the DRS.

For some tables (specifically, for higher frequency output), output is only requested for selected time slices of the experiments.

Independent assessment from DKRZ suggests that supplying data for all ensemble members and for the complete time periods would increase the archive size 10-fold.

As detailed below, it appears difficult to reconcile the definition of "requested" data implicit in the CMIP5 spreadsheet distributed to data suppliers with that implied by the DRS (version 27). In order to be clear about which "requested" is being referred to, I will use "CMIP5:requested" to refer to the request sent out to modelling groups and use "DRS:requested" or "product=requested" to refer to its usage in the DRS.

3 Archive size: replicated versus cmip5:requested

3.1 Variables omitted

The following collections of variables in the cmip5:requested output are omitted from the replicated subset:

3.1.1 Annual ocean data

frequency:annual/realm:ocean/variables:{low priority}

This reduces the number of ocean annual variables (all 3d) from 57 to 16.

3.1.2 Monthly ocean data

frequency:monthly/realm:ocean/variables:{3d, low and medium priority}

This reduces the number of 3d variables ocean monthly variables from 19 to 9.

3.1.3 6 hourly atmospheric data

expt:{historical, AMIP,RCP4.5, RCP8.5, AMIP-atmos}/frequency:6hr/realm:atmos/ variables:{all of table 6hrLev}

Note that 6 hourly pressure level data for these experiments is still to be replicated, so this partition cannot be represented using the DRS frequency category.

3.2 Reduced time coverage

The following collections of variables will have reduced time coverage:
realm:atmos/frequency:3hr/expt:{emission-historical}/variables:{in table 3hr}: reduced from 56 to 46 years.
This reduction in time coverage reduces the estimated size of the replicated portion by 1Tb. It would be a useful simplification to remove this change in time coverage.

3.3 Discussion

When the DRS version 27 was agreed on, there was agreement that replication should take place at the same level of granularity as publication on the ESG data node. Any finer granularity would be impossible to track in the catalogue. It was intended that the “product=requested” data collection would serve this purpose, and that the replicated data would be made up of a subset of the published units in the “product=requested” collection. It has been decided that data will be published at the “realm” level, meaning that a ESG publication unit (EPU) will be all data with a given: “<activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>” and version. Given this definition of a publication unit, the replicated data for monthly/ocean data will be a subsets of the CMIP5:requested data EPU.

I believe the decision to proceed with two product categories was based on a misunderstanding of the criteria specifying the replicated subset. In order to fit replication into subset of EPUs in the currently defined “product=requested” data collection, we would have to omit high priority and 2d ocean variables, as well as 6 hourly pressure level atmospheric data: this option appears too drastic. Therefore, we need a “product=replicated” element of the DRS.

Do we also need a “product=requested” branch? The “CMIP5:requested” variables and time periods are likely to be more widely available than others – could this information be useful? E.g. if a user wants all the requested monthly ocean data or all the requested 3 hourly data? These requests could be done by specifying a subset of experiments in the first case and a subset of experiments combined with a set of time slices in the second. These options will be available, so the only question is whether we want to provide a shorthand which makes accessing such a subset through wget trivial? Given the current deadlines, I'm inclined to say it isn't worth it.

4 Other DRS issues

4.1 *Data beyond the CMIP5 request*

As noted above, some modelling centres expect to archive (and process into CMOR2 compliant NetCDF) more data than specified in the CMIP5 request. This section raises a few questions about just how much additional data will be supported by level 1 QC process which data must pass to get into the ESGF CMIP5 archive. The most likely form of additional data will be extended time coverage and increased ensemble numbers, which will clearly be supported. Two other options are listed below.

4.1.1 Experiments and frequencies

Additional experiments and output frequencies will not be supported (I can't envisage any way of making such additions work).

4.1.2 Additional variables

Question: if a variable is not in the request, but is in the CF standard name list, can it be included in the CMIP5 archive? The DRS states that the "variable identifier" is composed from the variable short name and the name of the table in which it occurs, which suggests adding variables may be deprecated. On the other hand, if a variable has a well defined short name and can be assigned to a table, why not allow modelling centres to archive and advertise it as such?

There are two distinct cases here:
** variables which do not occur anywhere in the request
** variables which occur in, for instance, the monthly request but which a centre wants to add to higher frequency output.

The issue here is to have a clear statement of what is intended to be supported by the system.

4.2 *The DRS "variable identifier"*

As noted above, the DRS states that the "variable identifier" is composed from the variable short name and the name of the table in which it occurs. The DRS does not, however, say how the variable short name and table name will be combined, though it does imply that neither hyphen nor underscore will be used (the first is ruled out for variables, the second for all components). The hyphen is ruled out because of a desire to have variable names which can be used as Fortran variable names – this objective also rules out ":". How about camel-case, e.g. "tasAmon", "tasDa", "tas6hrplev", "tas6hrlev" (only works in a user friendly way if variable names never end in an integer)?

Alternatively, we add an exception to the permitted character restrictions: namely, that the variable identifier has exactly 1 underscore. e.g. "tas_amon", "tas_da", "tas_6hrplev", "tas_6hrlev". This is more human friendly, but less machine friendly.

A third alternative is to make the table name a separate DRS component, but combine the variable name and table name into a single attribute in the netcdf file and hence in the THREDDS catalogue. This would look like the second option to users, but would have the advantage of having a clearer (more uniform) specification of the different controlled vocabularies.