Data Request Review 2018

June 21st, Reading

Attending the meeting: Karl Taylor, Bryan Lawrence (in and out), Martin Juckes, Matthew Mizielinski, Alison Pamment, Charlotte Pascoe (afternoon), Emma Hogan, Antonio Cofino

Input from Stephane Senesi by email.

Summary of outcomes

In the near term

1. General support for maintaining the structure of the request, minimal changes, but there is a need to make the process run more smoothly, and to distribute work;
2. The data request has a critical function in maintaining consistency of data specifications and enabling greater automation in data delivery workflows;
3. The release process can be improved (benefitting from the stability given by a clear common understanding of the process and the links to co-dependent services): in future releases there will be a beta version made available for initial testing, followed by a full release in 3 weeks or sooner if given the go-ahead by by PCMDI and at least one other centre (the Hadley Centre volunteered to examine the beta version);
4. After PAMIP and CDRMIP have been added, we do not expect any further MIPs to join CMIP6 and, following a final revision of the experiment CV in July, we do not expect changes to the experiment list;
5. A new working group will be formed by the WIP to determine where improvements might be made and to oversee the evolution of the data request process -- the working group will be responsible for overseeing the scoping of the work, including communicating with the community (through the WIP where appropriate) about resource implications of any organisational decisions affecting the data request work; also monitoring and reporting on available resources and consequences of any shortfall;
6. The working group will organise a survey of the modelling groups and MIPs to gather feedback on the use of the data request;
7. Version/date should be recorded by the data request indicating when information was harvested from ES-DOC, CMIP6-CVs and CF Standard name list, and giving version information where possible (CF standard name version number, CMIP6 CV version stamp);
8. A baseline request should be added for all CMIP6 experiments, to ensure that key variables are not omitted; this is based on the group of variables defined for ESMVal[1] , but excluding the daily pressure level fields. Variables which are requested for either

[1] A list is available here: http://clipc-services.ceda.ac.uk/dreq/u/b7d445ce-c16a-11e6-bb6a-ac72891c3257.html

the historical or piControl simulation by 7 or more MIPs are also added to the request from ScenarioMIP experiments. Finally, all fixed fields, masks and area types are added to the request from all experiments.and an extended set for ScenarioMIP.

9. Attempt to gather information from modelling centres about the variables they intend to provide, and at what resolutions.

Beyond CMIP6

1. For CMIP7, and any individual MIPs that happen between CMIP6 and CMIP7, we need to rationalise the process of data request generation and maintenance, and, where possible, implement changes to simplify use of the data request;

2. MIPs outside CMIP would be supported (subject to funding) as independent projects in the sense that there would be no dependency on experiments from other MIPs within the request. This will make it possible to add MIPs as a simple extension to the existing request database;

3. Variable definitions, on the other hand, should be maintained as community definitions which can be re-used;

4. The “grid options” feature in the data request, which allows MIPs to specify a preference for the output grid is confusing and makes it difficult to aggregate requests; modelling centres do not have the capacity to respond to bespoke regridding requests, so the feature is of questionable use. This feature can be removed, and regridding options should be decided outside the data request. Other features which are little used and/or redundant will also be reviewed (taking into account the survey mentioned in point 6 above);

5. The way in which experiments are aggregated into groups is semantically awkward in the data request (requests can link directly to experiments, to a group of experiments or to all experiments in a MIP -- this approach evolved at an early stage of the request development and became frozen in to the system), this can be rationalised to simplify the use of the request;

6. Initial experiment information for CMIP6 was gathered directly from the MIP co-chairs: in future it will be taken from ES-DOC and “CMIP6 Cvs” (the Cvs govern the terminology to be used in ESGF, ES-DOC has more detailed information), now that these have stable scope and interfaces. Establishing this workflow will reduce duplication and make the purpose of information gathering clearer. It will mean that experiments do not enter the data request until initial information is entered in the upstream sources;

7. While preserving the structural integrity of the request as a coherent database, it will be helpful to split it into a small set of independently managed documents. For instance, a “variable definition section” could be split from the “request linking section” and an experiment specification section: the variable definitions are part of the community standards process, intended to be re-used across many MIPs; the request linking section can be split cleanly into parts associated with each MIP. This split would be transparent as far as the API is concerned. For many independent MIPs this request component will be trivial (one selection of variables for all experiments), the complexity mainly arises from the cross linkages of the many MIPs involved in CMIP6. Details to be discussed.

8. Tracking of discussions, including the interaction with the CF standard name approvals process will be improved: e.g. each MIP will be asked to submit a list of new variables in a spreadsheet (as done for CMIP6)[2]; each variable will be assigned a unique identifier and, after initial feedback, variables that need new names will be fed into the CF Standard Name editor, and email discussions will then be able to reference the editor pages. Github issues will also be used to monitor progress and collect information about associated CF metadata;

9. The use of vocabularies to organise information in the request and, later, in ESGF, will be reviewed to ease conflicts between early stability in the organisation of information being requested and the need to adjust some of the ESGF facets at a later stage;

10. The python library and the document should be separately versioned;

11. Review and extend the “baseline” request, ensure diagnostics needed for diagnostic packages (ESMVal and PCMDI Metrics Package) are covered.

[2] Other options may be possible, but the spreadsheet approach is supported by the current CF Standard name approvals process.