



The CMIP5 questionnaire: web-based metadata collection for climate modelling

Gerard Devine (1), Bryan Lawrence (2), Charlotte Pascoe (2), Rupert Ford (3), Paul Slavin (3), and Metafor Team ()

(1) NCAS-CMS, Department of Meteorology, University of Reading, United Kingdom (g.m.devine@reading.ac.uk), (2) Rutherton Appleton Laboratory, STFC, United Kingdom, (3) University of Manchester, United Kingdom

Simulations are currently underway in support of the next climate model intercomparison project (CMIP5). The sheer volume of data arising from these simulations presents an unprecedented technical challenge, both in terms of eventual data discoverability and the understanding of the provenance of that data. In response to this, the EU-funded METAFOR project has developed a web-based metadata collection tool, or questionnaire, which has now been released to the modelling community. The questionnaire is collecting metadata across the full spectrum of the climate modelling process, covering details of the climate models themselves, the grids and computing platforms on which the models were executed and the simulations they performed. The questionnaire captures specific details about how each simulation conformed to a particular CMIP5 experiment, as well as information on external data files.

Several technologies have been used successfully in the generation of the questionnaire, most notably Django, a python-based framework, that has allowed for efficient generation of the web interfaces, and a number of xml-based tools that have allowed a useful separation of the different stages in the questionnaire workflow. The information asked by the questionnaire originates from a pre-generated controlled vocabulary (CV), where the level of the information requested represents a balance between scientific usefulness and the human effort needed to complete it.

Upon completion of the questionnaire, the information is validated and then persisted in CIM-compliant xml documents, where CIM is the 'Common Information Model' for climate science. This standardised format allows the information entered via the questionnaire to be harvested by and thus discoverable through CIM-aware web portals, for example the Earth System Grid (ESG). This opens up a variety of services and tools that can operate upon the harvested metadata and, in the first instance, will allow for a range of search and differencing functionalities.

The CMIP5 set of experiments represents a huge scientific effort and the metadata questionnaire will lead to an invaluable set of metadata on which to explore CMIP5 science. However, the questionnaire itself is not explicitly engineered for CMIP5. Generation of a new CV to describe statistical downscaling for regional climate analysis is currently underway and will illustrate the extensibility of the questionnaire, in this case for a major source of climate information.