Geophysical Research Abstracts
Vol. 17, EGU2015-9445, 2015
EGU General Assembly 2015

Versioning for CMIP6 in the Earth System Grid Federation

Tobias Weigel (1,2), Stephan Kindermann (1), and Michael Lautenschlager (1)

(1) German Climate Computing Center, Department of Data Management, Hamburg, Germany (weigel@dkrz.de), (2) Universität Hamburg, Germany

The Earth System Grid Federation (ESGF) has been used as the e-infrastructure to provide access to CMIP5 data and is expected to serve CMIP6 data as well. 2015 marks the year of continued planning and preparation where new concepts can still be implemented for the operational phase of CMIP6.

A particular concern within ESGF operations is the versioning and automated replication of data. From CMIP5 experience we know that the pathway between initial submission of modelling data to the ESGF data space and quality-controlled long-term archival of the final products is long and far from linear. Data may be retracted, amended and updated, and metadata may accumulate at different stages. It is unrealistic to assume that a simple and straightforward process can be used as a role model to build ESGF services around the different stages data will pass through during the active phase of CMIP6. Nonetheless, at the technical level ESGF requires some form of automated control and management. At the same time, the accountability of data products must be made transparent to guard against misinterpretation, increase user experience and promote open and reproducible science.

To address the challenges, first some essential versioning policies must be agreed upon and enforced through technical means and organizational processes. The volatile readiness state of CMIP data cannot be changed as it is given by the users; however its management can be improved. A promising approach is to embed persistent identifiers in all CMIP6 data objects and register them so they can be globally resolved by any user and used as reference points within ESGF management processes. A specific conceptual interpretation and management of such identifiers can ensure that they remain valid and useful even if the data objects change or become unavailable. For this, identifiers must be assigned to individual versions and aggregations, connected with each other and integrated in the existing ESGF publication process. This can also facilitate automatic replication procedures necessary in view of the expected CMIP6 data volumes. Identifier management will require stable RESTful APIs and associated tools for CMIP6 operations. The resulting framework can have benefits beyond core ESGF use cases including applications in other disciplinary infrastructures.

Some initial prototyping has already been done and the newly formed ESGF working team on replication and versioning will follow a dedicated roadmap to prepare the e-infrastructure for the upcoming challenge.