Geophysical Research Abstracts
Vol. 18, EGU2016-13297, 2016
EGU General Assembly 2016
© Author(s) 2016. CC Attribution 3.0 License.

Persistent identifiers for CMIP6 data in the Earth System Grid Federation

Merret Buurman (1), Tobias Weigel (1), Martin Juckes (2), Michael Lautenschlager (1), and Stephan Kindermann (1)

(1) German Climate Computing Centre (DKRZ), Hamburg, Germany (buurman@dkrz.de), (2) STFC, BADC, Chilton, Didcot, Oxon, United Kingdom

The Earth System Grid Federation (ESGF) is a distributed data infrastructure that will provide access to the CMIP6 experiment data. The data consist of thousands of datasets composed of millions of files. Over the course of the CMIP6 operational phase, datasets may be retracted and replaced by newer versions that consist of completely or partly new files. Each dataset is hosted at a single data centre, but can have one or several backups (replicas) at other data centres.

To keep track of the different data entities and relationships between them, to ensure their consistency and improve exchange of information about them, Persistent Identifiers (PIDs) are used. These are unique identifiers that are registered at a globally accessible server, along with some metadata (the PID record). While usually providing access to the data object they refer to, as long as it exists, the metadata record will remain available even beyond the object's lifetime. Besides providing access to data and metadata, PIDs will allow scientists to communicate effectively and on a fine granularity about CMIP6 data. The initiative to introduce PIDs in the ESGF infrastructure has been described and agreed upon through a series of white papers governed by the WGCM Infrastructure Panel (WIP).

In CMIP6, each dataset and each file is assigned a PID that keeps track of the data object's physical copies throughout the object lifetime. In addition to this, its relationship with other data objects is stored in the PID recordA human-readable version of this information is available on an information page also linked in the PID record. A possible application that exploits the information available from the PID records is a smart information tool, which a scientific user can call to find out if his/her version was replaced by a new one, to view and browse the related datasets and files, and to get access to the various copies or to additional metadata on a dedicated website.

The PID registration process is embedded in the ESGF data publication process. During their first publication, the PID records are populated with metadata including the parent dataset(s), other existing versions and physical location. Every subsequent publication, un-publication or replica publication of a dataset or file then updates the PID records to keep track of changing physical locations of the data (or lack thereof) and of reported errors in the data.

Assembling the metadata records and registering the PIDs on a central server is a potential performance bottleneck as millions of data objects may be published in a short timeframe when the CMIP6 experiment phase begins. For this reason, the PID registration and metadata update tasks are pushed to a message queueing system facilitating high availability and scalability and then processed asynchronously. This will lead to a slight delay in PID registration but will avoid blocking resources at the data centres and slowing down the publication of the data so eagerly awaited by the scientists.