Geophysical Research Abstracts
Vol. 14, EGU2012-9124, 2012
EGU General Assembly 2012

Web Services and Handle Infrastructure – WDCC's Contributions to International Projects

G. Föll, T. Weigelt, S. Kindermann, M. Lautenschlager, and F. Toussaint
Deutsches Klimarechenzentrum, World Data Centre for Climate, Hamburg, Germany (toussaint@dkrz.de)

Climate science demands on data management are growing rapidly as climate models grow in the precision with which they depict spatial structures and in the completeness with which they describe a vast range of physical processes. The ExArch project is exploring the challenges of developing a software management infrastructure which will scale to the multi-exabyte archives of climate data which are likely to be crucial to major policy decisions in by the end of the decade.

The ExArch approach to future integration of exascale climate archives is based on one hand on a distributed web service architecture providing data analysis and quality control functionality across archvies. On the other hand a consistent persistent identifier infrastructure is deployed to support distributed data management and data replication.

Distributed data analysis functionality is based on the CDO climate data operators' package. The CDO-Tool is used for processing of the archived data and metadata. CDO is a collection of command line Operators to manipulate and analyse Climate and forecast model Data. A range of formats is supported and over 500 operators are provided. CDO presently is designed to work in a scripting environment with local files. ExArch will extend the tool to support efficient usage in an exascale archive with distributed data and computational resources by providing flexible scheduling capabilities.

Quality control will become increasingly important in an exascale computing context. Researchers will be dealing with millions of data files from multiple sources and will need to know whether the files satisfy a range of basic quality criterea. Hence ExArch will provide a flexible and extensible quality control system. The data will be held at more than 30 computing centres and data archives around the world, but for users it will appear as a single archive due to a standardized ExArch Web Processing Service.

Data infrastructures such as the one built by ExArch can greatly benefit from assigning persistent identifiers (PIDs) to the main entities, such as data and metadata records. A PID should then not only consist of a globally unique identifier, but also support built-in facilities to relate PIDs to each other, to build multi-hierarchical virtual collections and to enable attaching basic metadata directly to PIDs. With such a toolset, PIDs can support crucial data management tasks. For example, data replication performed in ExArch can be supported through PIDs as they can help to establish durable links between identical copies. By linking derivative data objects together, their provenance can be traced with a level of detail and reliability currently unavailable in the Earth system modelling domain. Regarding data transfers, virtual collections of PIDs may be used to package data prior to transmission. If the PID of such a collection is used as the primary key in data transfers, safety of transfer and traceability of data objects across repositories increases. End-users can benefit from PIDs as well since they make data discovery independent from particular storage sites and enable user-friendly communication about primary research objects. A generic PID system can in fact be a fundamental building block for scientific e-infrastructures across projects and domains.