Geophysical Research Abstracts
Vol. 15, EGU2013-7240, 2013
EGU General Assembly 2013

EUDAT: A New Cross-Disciplinary Data Infrastructure For Science

Damien Lecarpentier (1), Alberto Michelini (2), and Peter Wittenburg (3)

(1) CSC, IT Center for Science, Finland (Damien.Lecarpentier@csc.fi), (2) Istituto Nazionale di Geofisica e Vulcanologia, Italy (alberto.michelini@ingv.it), (3) Max Planck Institute for Psycholinguistics, The Netherlands(Peter.Wittenburg@mpi.nl)

In recent years significant investments have been made by the European Commission and European member states to create a pan-European e-Infrastructure supporting multiple research communities. As a result, a European e-Infrastructure ecosystem is currently taking shape, with communication networks, distributed grids and HPC facilities providing European researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. However, the accelerated proliferation of data – newly available from powerful new scientific instruments, simulations and the digitization of existing resources – has created a new impetus for increasing efforts and investments in order to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation.

EUDAT is a pan-European initiative that started in October 2011 and which aims to help overcome these challenges by laying out the foundations of a Collaborative Data Infrastructure (CDI) in which centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities. Although research communities from different disciplines have different ambitions and approaches – particularly with respect to data organization and content – they also share many basic service requirements. This commonality makes it possible for EUDAT to establish common data services, designed to support multiple research communities, as part of this CDI.

During the first year, EUDAT has been reviewing the approaches and requirements of a first subset of communities from linguistics (CLARIN), solid earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), and shortlisted four generic services to be deployed as shared services on the EUDAT infrastructure. These services are data replication from site to site, data staging to compute facilities, metadata, and easy storage. A number of enabling services such as distributed authentication and authorization, persistent identifiers, hosting of services, workspaces and centre registry were also discussed.

The services being designed in EUDAT will thus be of interest to a broad range of communities that lack their own robust data infrastructures, or that are simply looking for additional storage and/or computing capacities to better access, use, re-use, and preserve their data. The first pilots were completed in 2012 and a pre-production ready operational infrastructure, comprised of five sites (RZG, CINECA, SARA, CSC, FZJ), offering 480TB of online storage and 4PB of near-line (tape) storage, initially serving four user communities (ENES, EPOS, CLARIN, VPH) was established. These services shall be available to all communities in a production environment by 2014.

Although EUDAT has initially focused on a subset of research communities, it aims to engage with other communities interested in adapting their solutions or contributing to the design of the infrastructure. Discussions with other research communities – belonging to the fields of environmental sciences, biomedical science, physics, social sciences and humanities – have already begun and are following a pattern similar to the one we adopted with the initial communities. The next step will consist of integrating representatives from these communities into the existing pilots and task forces so as to include them in the process of designing the services and, ultimately, shaping the future CDI.