Contribution ID: **14** Type: **Tutorial**

# A Tutorial on Hybrid Data Infrastructures: D4Science as a case study

*Thursday, 12 November 2015 09:00 (1h 30m)*

An e-Infrastructure is a distributed network of service nodes, residing on multiple sites and managed by one or more organizations allowing scientists residing at distant places to collaborate. They may offer a multiplicity of facilities as-a-service, supporting data sharing and usage at different levels of abstraction. E-Infrastructures can have different implementations (Andronico et al 2011). A major distinction is between (i) Data e-Infrastructures, i.e. digital infrastructures promoting data sharing and consumption to a community of practice (e.g. MyOcean, Blanc 2008) and (ii) Computational e-Infrastructures, which support the processes required by a community of practice using GRID and Cloud computing facilities (e.g. Candela et al. 2013). A more recent type of e-Infrastructure is the Hybrid Data Infrastructure (HDI) (Candela et al. 2010), i.e. a Data and Computational e-Infrastructure that adopts a delivery model for data management, in which computing, storage, data and software are made available as-a-Service. HDIs support, for example, data transfer, data harmonization and data processing workflows. Hybrid Data e-Infrastructures have already been used in several European and international projects (e.g. i-Marine 2011; EuBrazil OpenBio 2011) and their exploitation is growing fast supporting new projects and initiatives, e.g. Parthenos, Ariadne, Descramble.
A particular HDI, named D4Science (Candela et al. 2009), has been used by communities of practice in the fields of biodiversity conservation, geothermal energy monitoring, fisheries management, and culture heritage. This e-Infrastructure hosts models and resources by several international organizations involved in these fields. Its capabilities help scientists to access and manage data, reuse data and models, obtain results in short time and share these results with other colleagues. In this tutorial, we will give an overview of the D4Science capabilities; in particular, we will show practices and methods that large international organizations like FAO and UNESCO apply by means of D4Science. At the same time, we will explain how the D4Science facilities conform to the concepts of e-Infrastructures, Virtual Research Environments (VREs), data sharing and experiments reproducibility. In our tutorial, we will give insight about how D4Science contributors can add new models and algorithms to the processing platform. D4Science adopts methods to embed software developed by communities of practice involving people with limited expertise in Computer Science. Community software involves legacy programs (e.g. written in Fortran 90) as well as R scripts developed under different Operating Systems and versions of the R interpreters. D4Science is able to manage this multi-language scenario in its Cloud computing platform (Coro et al. 2014). Finally, D4Science uses the EGI Federated Cloud (FedCloud) infrastructure for data processing: computations are parallelized by dividing the input in several chunks and each chunk is sent to D4Science services residing on FedCloud (Generic Workers) to be processed. Furthermore, another D4Science service executing data mining algorithms (DataMiner) also resides on FedCloud and adopts an interface that is compliant with the Web Processing Service (WPS, Schut and Whiteside 2015) specifications.

## Links, references, publications, etc.

Blanc, F. 2008. "MyOcean information system."In Proceedings of EuroGOOS 2008
Candela, L., Castelli D., and Pagano P. 2009. "D4Science: an e-Infrastructure for Supporting Virtual Research Environments." In Proceedings of IRCDL 2009
Candela, L., Castelli, D., Coro, G., Pagano, P., and Sinibaldi, F. 2013. "Species distribution modeling in the cloud."Concurrency and Computation: Practice and Experience. doi: 10.1002/cpe.3030
Candela, L., Castelli D., and Pagano P. 2010. "Making Virtual Research Environments in the Cloud a Reality: the gCube Approach." ERCIM News 2010.83: 32.
Coro, G., Candela, L., Pagano, P., Italiano, A., and Liccardo L. 2014. "Parallelizing the execution of native data mining algorithms for computational biology." Concurrency and Computation: Practice and Experience doi: 10.1002/cpe.3435
EuBrazil OpenBio. 2011. "The EuBrazil OpenBio Project."URL http://www.eubrazilopenbio.eu
EUDAT. 2015. "The EUDAT web site."URL eudat.eu
i-Marine. 2011. "The i-Marine European Project."URL http://www.i-marine.eu
Schut, P. and Whiteside, A. 2007. "OpenGIS Web Processing Service". OGC project document.

G. Andronico et al. (2011) e-Infrastructures for e-Science: A Global View. Journal of Grid Computing, 9(2), 155-184, doi: 10.1007/s10723-011-9187-y

## Additional information

This session intends to provide all information to learn how to run existing models on distributed resources while preserving control of access and confidentiality. It will explain how simple is
- the creation of a personal environment, named Virtual Research Environment
- the management of authorization policies: who can enter the VRE, who is allowed to use a service, etc.
- the execution of either existing or new models
- the monitoring of the execution
- the re-execution of a model and the comparison of the results
- the sharing of the results

Links
The D4Science Web Site: www.d4science.org
The D4Science Services Web Portal: services.d4science.org
The i-Marine Services Web Portal: i-marine.d4science.org
The gCube software Web Site: www.gcube-system.org

**Primary authors:** PAGANO, Pasquale (CNR); CORO, gianpaolo (CERN)

**Co-author:** CASTELLI, Donatella (Consiglio Nazionele delle Ricerche (CNR) - ISTI)

**Presenters:** PAGANO, Pasquale (CNR); CORO, gianpaolo (CERN)

**Session Classification:** Tutorial: Data and Processes without Boundaries: D4Science as a case study