An EMSO data case study within the INDIGO-DC project

Stephen Monna (1), Nicola M. Marcucci (1), Giuditta Marinaro (1), Sandro Fiore (2), Alessandro D'Anca (2), Marica Antonacci (3), Laura Beranzoli (1,4), Paolo Favali (1,4)

(1) INGV, Roma2, Rome, Italy (stephen.monna@ingv.it), (2) CMCC Foundation (Italy), (3) INFN (Italy), (4) EMSO Interim Support Team

We present our experience based on a case study within the INDIGO-DataCloud (INtegrating Distributed data Infrastructures for Global ExplOitation) project (www.indigo-datacloud.eu). The aim of INDIGO-DC is to develop a data and computing platform targeting scientific communities.

Our case study is an example of activities performed by INGV using data from seafloor observatories that are nodes of the infrastructure EMSO (European Multidisciplinary Seafloor and water column Observatory)-ERIC (www.emso-eu.org). EMSO is composed of several deep-seafloor and water column observatories, deployed at key sites in the European waters, thus forming a widely distributed pan-European infrastructure. In our case study we consider data collected by the NEMO-SN1 observatory, one of the EMSO nodes used for geohazard monitoring, located in the Western Ionian Sea in proximity of Etna volcano.

Starting from the case study, through an agile approach, we defined some requirements for INDIGO developers, and tested some of the proposed INDIGO solutions that are of interest for our research community.

Given that EMSO is a distributed infrastructure, we are interested in INDIGO solutions that allow access to distributed data storage. Access should be both user-oriented and machine-oriented, and with the use of a common identity and access system. For this purpose, we have been testing:

- ONEDATA (https://onedata.org), as global data management system.

- INDIGO-IAM as Identity and Access Management system.

Another aspect we are interested in is the efficient data processing, and we have focused on two types of INDIGO products:

- Ophidia (http://ophidia.cmcc.it), a big data analytics framework for eScience for the analysis of multidimensional data.

- A collection of INDIGO Services to run processes for scientific computing through the INDIGO Orchestrator.