OpenBioMaps data management service for biological sciences and biodiversity conservation

PI: Miklós Bán
Shepherd: Miguel Caballer

eosc-hub.eu

@EOSC_eu

EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536.

Overview

- The OpenBioMaps (OBM) is a free and open-source database framework for scientific and conservation purposes.
- Science areas:
  - Biology, conservation biology, ecology, biodiversity
- The OBM system is used for data management by nature conservation institutes, biodiversity research and citizen science projects.
- OBM provides a number of services that make day-to-day work with data easier, but it does not yet provide tools for analyzing the data. Particularly for high computing tasks.
- Current use cases:
  - https://openbiomaps.org/projects/

OpenBioMaps
spatial biotic databases

Overview

- Impressions of some projects

OpenBioMaps
spatial biotic databases

Main objectives

- Develop a background service that supports the interpretation of data from databases on conservation biology and biodiversity.
- A solution that facilitates and generalizes the most common high-computational analysis of data stored in such databases.
- We would like to create a service with EOSC that allows multiple users to run tasks that are above the level of a PC through the same interface. In fact, we would like to develop a “service in service” - specifically for projects that collect nature conservation and biodiversity data.

Computing tasks

- The most common computing tasks come from the following areas:
  - Spatial analyses (satellite image processing, large or fine scale spatial queries),
  - ML analyses (satellite image analyses, drone image analyses, distribution models (Random Forest), population dynamics analyses, survival analyses, supervised and non-supervised image classification - photos of individual animals, habitats, survey methods),
  - Conservation genetics analyses.
- To serve these diverse tasks we need a fully configurable VM which let us deploy our service interface (API) which will be available in the OpenBioMaps Network and provide computation capacity access to the involved projects.

EOSC-hub

Computing requirements

- According to our recent experiences in our PC based local computational cluster, the number of processors is the most important in these ecological analyzes. A “typical” analysis is now running at an acceptable (few hours to few days) rate on **16 CORES**.
- The parallel computing requirements of image analysis can be much higher, and **GPU usage can be interesting** there. Some analyzes, for example, genetic analyzes or larger spatial analyzes require **a lot of memory**.

Computing requirements

- Minimal Compute and Storage capacity needed for sustaining the Project:
  - 32 CPU cores
  - 96 Gb RAM (3Gb RAM per cores)
  - 2 Tb HD
- Compute and Storage capacity to fully scale-up the Project after the completion of the pilot:
  - 128 CPU cores
  - 384 Gb RAM (3Gb RAM per cores)
  - 8 Tb HD
- GPU access will be interesting.
- **IFCA site** has accepted to provide resources to this EAP case.

Application Goals in first year

1. Developing a new and flexible API interface together with the OBM user community for supporting conservation science.
2. Creating a database-based machine learning analysis system that can significantly support the use of data collected in nature conservation and improve the way new data is collected.
3. Creating a long-term computational layer for OBM

Work Plan I.

- Q1:
  - Integration with EGI Cloud Compute.
  - Deploy and Configure OBM node in test environment manually.
- Q2:
  - Create TOSCA Recipes and Ansible roles needed to deploy the application automatically using IM.

Work Plan II.

- Q3:
  - Deploy OBM node to production environment using the developed recipes.
  - Analyse EOSC data services to be used by the application:
    - EGI DataHub.
    - B2 services (Drop, Find, Handle, Share)
  - EGI Services (Training Infrastructure, Data Transfer)
  - EOSC Marketplace services (GeoDAB, D4Science spatial services, Alien and Invasive Species Virtual Research Environment, Biodiversity, EODC JupyterHub for global Copernicus data)
- Q4:
  - Performance test of all nodes.

Budget

- Person Months: 3,5.
- Budget: 21.658 €
  - 3 PMs to IFCA as resource provider.
    - One VM with 48 vCPU, 96GB RAM, 2 Tb HD
  - 0,5 PM to UPV for integration issues.

Thank you for your attention!

Questions?

Contact

banm@vocs.unideb.hu
micafer1@upv.es

eosc-hub.eu @EOSC_eu