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 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.
  - One VM with 48 vCPU, 96GB RAM, 2 Tb HD

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.
- 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.

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

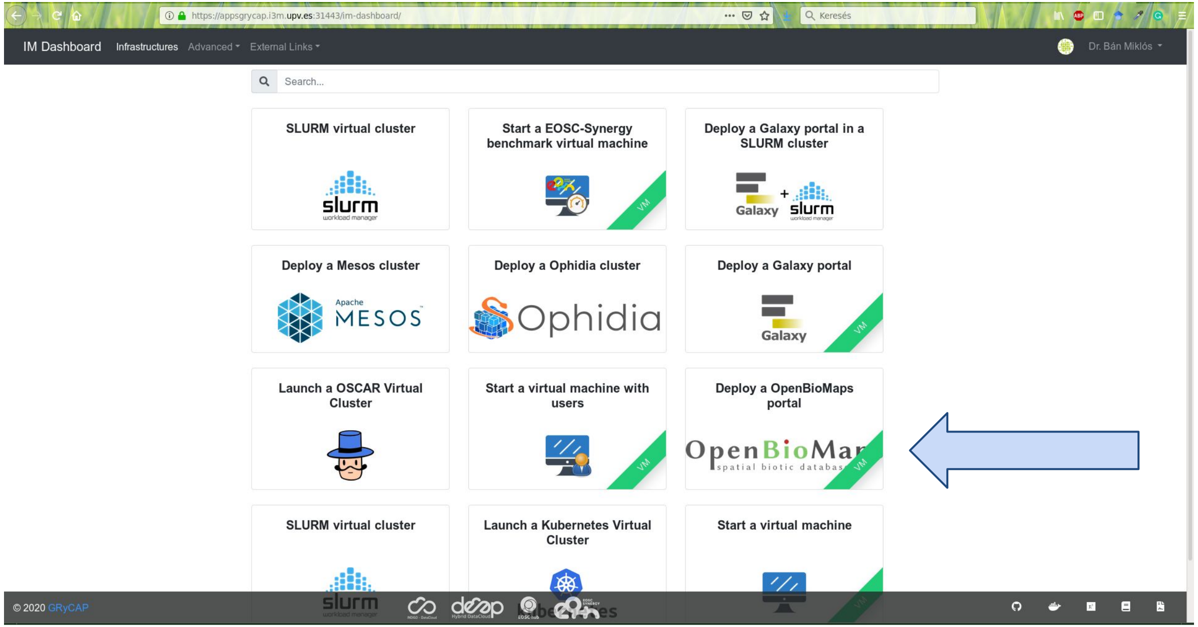
- The OpenBioMaps (OBM) is a free and open-source database framework for scientific and conservation purposes.
- Science areas:
  - Biology, conservation biology, ecology, biodiversity
- The OBM system is used for data management by nature conservation institutes, biodiversity research and citizen science 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/

Current status

- Status of the Wiki page (Created, filling in progress):
  - https://wiki.eosc-hub.eu/display/EOSC/OpenBioMaps+data+management+service+for+biological+sciences+and+biodiversity+conservation+pilot
- Status of the community requirement DB entry (Created, but not accessible for me):
    - https://wiki.eosc-hub.eu/display/EOSC/OpenBioMaps+data+management+service+for+biological+sciences+and+biodiversity+conservation
- Status of access to the services and SLAs agreed with the providers (Created):
  - SLA agreed with IFCA site (https://documents.egi.eu/public/ShowDocument?docid=3565)
    - 1-8 VM, 48 CPUs, 100 GB RAM, 2 TB HD. (It's still not clear to me, the size of the allocated resources, because it does not match my two agreements…)
- Progress of the application:
  - OBM computation layer: **in development**
  - OBM as EOSC service: **test node successfully deployed** at IFCA manually with OpenStack interface.
    - Initial version of TOSCA Recipes and Ansible roles created to deploy the application automatically using IM.
    - Currently testing the automated deployment using Infrastructure Manager Dashboard: https://appsgrycap.i3m.upv.es:31443/im-dashboard/

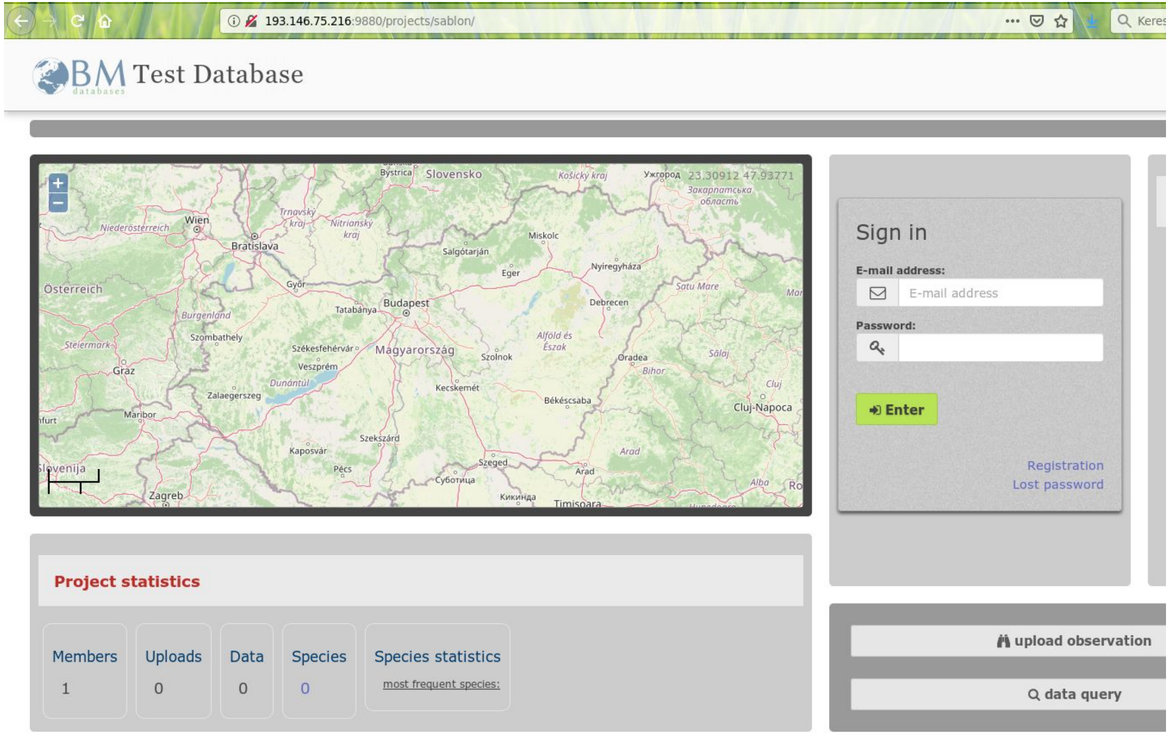
Current Status

- Created (with Miguel) to launch OpenBioMaps portal from IM dashboard



Current Status

- OpenBiomaps deployed Successfully!



193.146.75.216:9880/projects/sablon/

BM Test Database

Sign in

E-mail address: E-mail address

Password:

Enter

Registration

Lost password

Project statistics

Members	Uploads	Data	Species	Species statistics
1	0	0	0	most frequent species:

Background computational service for OpenBioMaps

- Creating background computational service for OpenBioMaps projects to help researcher using OBM databases to apply particular computations.
- What are these computations? Species distribution modelling using environmental information.

Why do we need a dedicated computing capacity?

- Because these are complex and complicated computations and because the data is mostly not generated where there would be capacity to process it. Anyone who would be able to perform these computations mostly doesn’t even know about the data or doesn’t have access to it.
  - This is because much of the biodiversity data is collected by conservationists and amateurs
- Solution: If I can create a database background service that can be used to analyze data from the database without moving it (sharing, downloading or exporting), it is likely that a much larger proportion of biodiversity data will be processed and much more research collaboration will be created with conservation.

I have two early adopters in my EOSC-EAP project

both use OpenBioMaps, but in different contexts

- A research group that performs analysis of a large number of environment factors on the state variables of mid-mountain forests. There are more than 200 environmental variables and distribution data of living organisms collected by different methods.
- A National Park which would like to perform automatic analysis of the spread of invasive plants from aerial photographs (drone and satellite images). Geoss service will be needed…

- Both cases, we are using ML algorithms in analyses

In both cases we have good result in the analysis part

(it is good news for me :D but maybe less interesting here.)

The interface between the OBM database server and the EOSC computational server is still under construction and cannot be published yet. I am currently busy with OBM side developments:

- **Service access management** (who and how can access computation capacity) - it is not ready yet
- **Data management interface** (which data selected and how send it the computational node) - it mostly works, but
  - Problem with large file transfers: integrating B2.. services. I don't know yet, how
  - Sharing data: B2Handle, B2Share - I will get help from our library to learn them..
  - Using Geoss / Geo DAB - not yet tested
- **Computational package interface** - will be an interface to create/share/upload custom computational packages - we haven't started creating yet.

Computational node's side developments:

- Request processing and management: We have not yet begun to accomplish these tasks
- Generalized computation packages: We have good results

Issues

- At first, things don’t work
- There is a lot of and diverse information, but it is often superficial. At first, the options always seem scary, but later when I search for the specifics, I don’t find enough information.
- E.g. on openstack interface:
  - Volumes cannot be attached to VMs
  - Orphaned volumes cannot be deleted

Thank you for your attention!

Questions?

Contact

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

EOSC-hub

eosc-hub.eu @EOSC_eu