GLIM

BioData.pt

Ready for BioData.pt Management?

EMO-BON and its associated data resources for end-users.

Management of biodiversity/Life data

David Paleček, Cymon J. Cox, Frederico Mestre, Andrzej Tkazc - Centro de Ciências do Mar (Faro, Portugal), **Katrina Exter, Marc Portier, Cedric Decruw, Laurian Van Maldeghem, Bram Ulrichts** - Flanders Marine Institute (Ostend, Belgium), **Stelios Ninidakis** - Hellenic Centre for Marine Research (Heraklion, Greece), **Maria Luisa Chiusano, Maria Chiara Langella, Ilia Mauriello** - University Federico II (Naples, Italy), **Marco Miralto** - Stazione Zoologica A. Dohrn (Naples, Italy), & **Ioulia Santi** - EMBRC HQ (Paris, France)

3rd November 2025

eosc | FAIR-EASE Building Interoperable Earth Science & Environmental Services

BioData.pt | elixir PORTUGAL

FCCN serviços digitais fct | fct Fundação para a Ciência e a Tecnologia | PRR Plano de Recuperação e Resiliência | REPÚBLICA PORTUGUESA | Financiado pela União Europeia NextGenerationEU

Outline

EMO BON
- sampling
- analyses

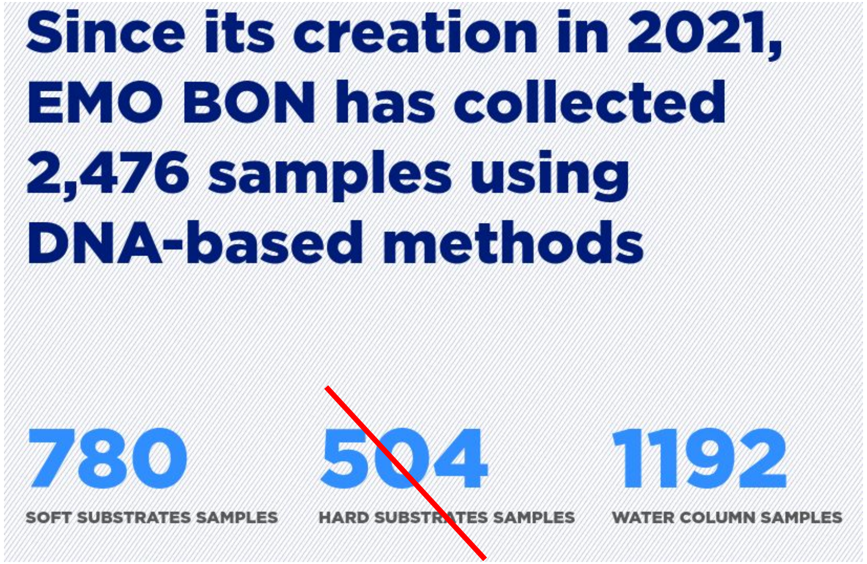
Data management and curation
- scientist’s expectations
- ideal Virtual Research Environment
- current status of the tools

Demonstration
- alpha and beta diversities
- taxonomy finder
- complex microbial networks

EMBRC and EMO-BON: sampling

EMO BON

- environmental DNA sampling + measurements of physical params including EOVs.
- metagenomics samples are automatically queued to EMBL MGnify.
- **standardisation** (SOPs), standard analyses, maximally comparable data sets, open, FAIR, and **interoperable** data stream
- 2000 samples taken, 1000 sequenced, 250 analysed

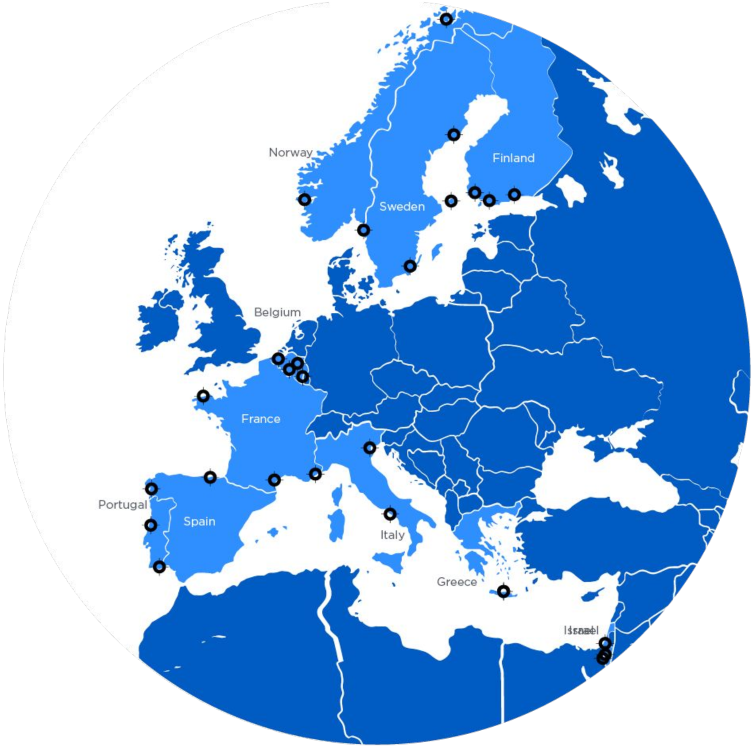


Since its creation in 2021, EMO BON has collected 2,476 samples using DNA-based methods

780	~~504~~	1192
SOFT SUBSTRATES SAMPLES	~~HARD SUBSTRATES SAMPLES~~	WATER COLUMN SAMPLES

EMBRC and EMO-BON: sampling

- sampling once every 2 months - 19 sites
- filters (3 um / 0.2 um) + sediments, published SOPs
- environmental sampling data contain 139 provenance and environmental metadata values recorded in Google logsheets (manual entry)
  - validated by python validators (*pydantic*)
  - quality control (CD/CI Github actions)
- EMBL BioSamples/BioProjects for observatory and project metadata
- samples sent from marine stations to EMBRC-HQ, Paris, and then shipped to Genoscope for sequencing in 6-monthly batches, ENA PRJEB51688
- duplicate samples are biobanked

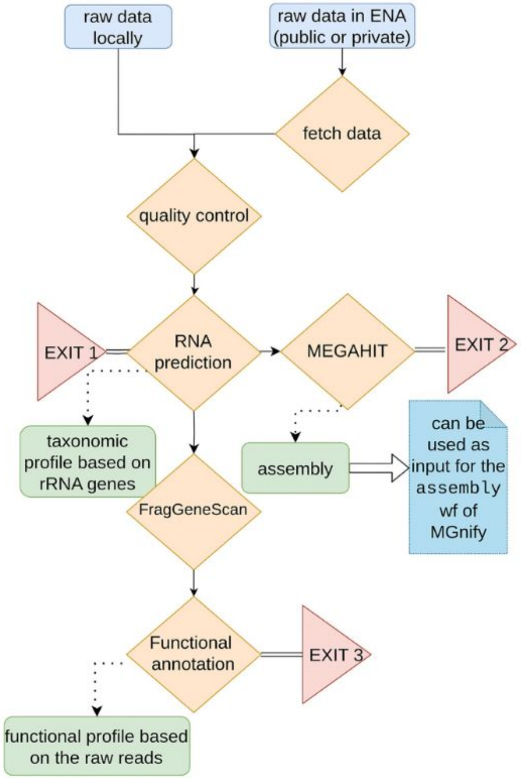


https://www.embrc.eu/emo-bon/

FCCN fct PRR Plano de Recuperação e Resiliência REPÚBLICA PORTUGUESA Financiado pela União Europeia NextGenerationEU

EMBRC and EMO-BON: analysis

- metaGOflow (EOSC-Life) workflow based on EMBL MGnify (but faster and not as extensive - no “assembled” or “systems biology” workflows) only uses the “reads workflow”
- 3 types of data products:
  - taxonomic inventories (one each based on **SSU** and **LSU** RNA molecules)
  - functional annotations (**InterProScan** - Gene Ontology, KEGG, and Pfam labels)
  - assembled sequences for downstream analyses
- single sample takes **3-8 days** depending on resources and complexity of the data - input read sequences (150 bp) range from **3-6GB** (forward and reverse Illumina read data)
- output is between **30-60GB** raw data (8-25GB compressed) - between **60-70 separate data files**
- *currently ~200 analysed using MGF (~5TB compressed data, 181 published), 1800 to go*


raw data locally
raw data in ENA (public or private)
fetch data
quality control
EXIT 1
RNA prediction
MEGAHIT
EXIT 2
taxonomic profile based on rRNA genes
assembly
can be used as input for the assembly wf of MGnify
FragGeneScan
Functional annotation
EXIT 3
functional profile based on the raw reads

H Zafeiropoulos et al., GigaScience, Volume 12, 2023, giad078

Ideal world: Virtual Research Environment

User requirements:

- Browse data as a web

  ontology using vocabularies, https://data.emobon.embrc.eu/ns ✅
- Pull additional data (EOVs) or other metagenomic campaign data (e.g. MGnify)

  RDF triples and RO-Crates ☑️
- Subsample/Filter data

  UDAL queries ➖ **SPARQL endpoint** ❌
- Access to a starting point visualizations

  Data Analysis Kit (Jupyter) ✅
- Analyze seamlessly from single environment

  Blue-Cloud deployment ❌
- Possibility to submit jobs to a cluster

  Galaxy access ☑️

eosc | Blue-Cloud2026

A federated European FAIR and Open Research Ecosystem
for oceans, seas, coastal and inland waters

GLIM BioData.pt | elixir PORTUGAL

Ready for
BioData.pt
Management?

FCCN | fct PRR Plano de Recuperação e Resiliência REPÚBLICA PORTUGUESA Financiado pela União Europeia NextGenerationEU

Ontology

What are we solving: findable, interoperable

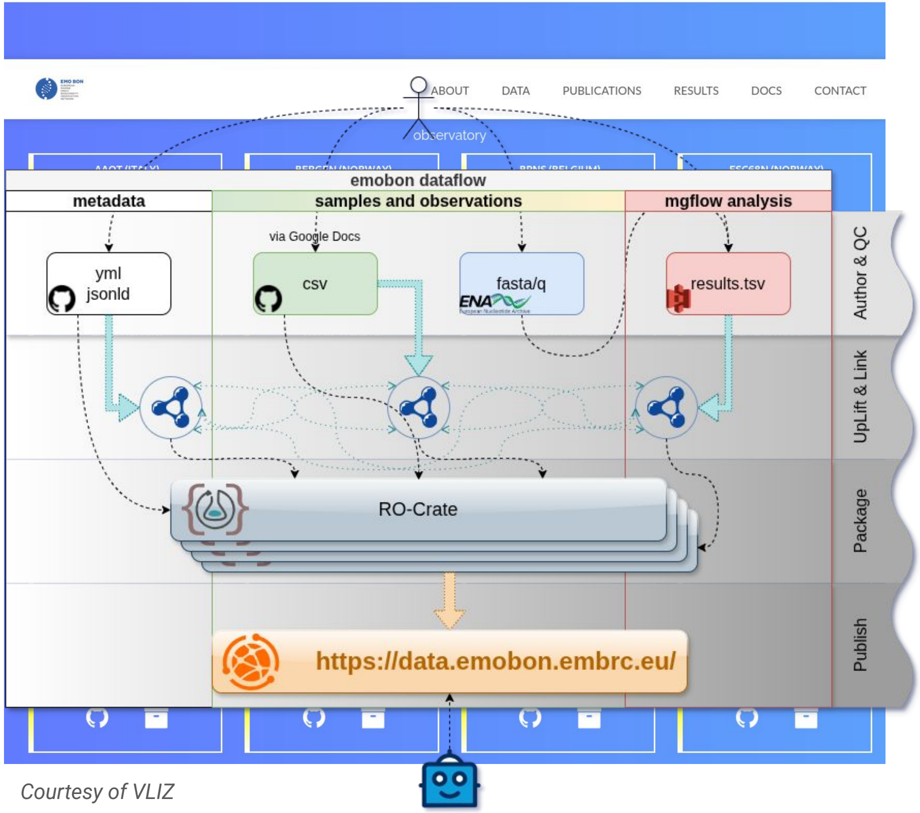
- defines the relations and properties of the data for the interoperability, provenance as triples
- unofficial, but useful in terms of vocabulary: https://github.com/emo-bon/observatory-profile/blob/main/logsheet_schema_extended.csv
- new ontology terms: https://data.emobon.embrc.eu/ns/

GLIM BioData.pt BioData.pt elixir PORTUGAL Ready for BioData.pt Management? FCCN fct PRR Plano de Recuperação e Resiliência REPÚBLICA PORTUGUESA Financiado pela União Europeia NextGenerationEU

RDF triples and RO-Crates: ☑

What are we solving: Data distribution, semantic search

- Python script builds the RO-crates: uploads to S3 store using DVC, does the semantic uplift of both the taxonomic inventories and functional annotation, and writes metadata.json file for the 60-70 file payload:
- RO-Crate viewer, https://data.emobon.embrc.eu/
- Single entry point to diverse interrelated data



Courtesy of VLIZ

GLIM BioData.pt
BioData.pt
elixir PORTUGAL
Ready for BioData.pt Management?
FCCN | fct
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU

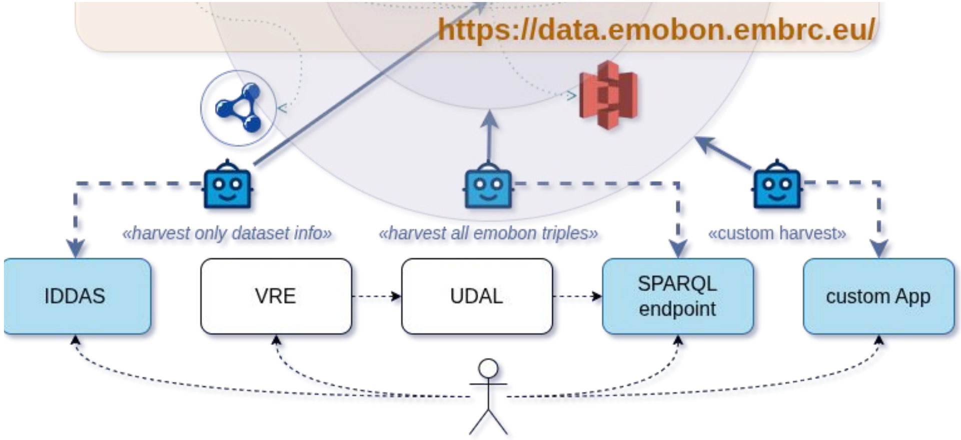
SPARQL

What are we solving: Public facing machine readable knowledge graph for SQL type queries

- Harvests links to datasets from data.emobon.embrc.eu/
- Assemble ALL linked triples (including data turtle inside RO-Crates, *yes nesting is allowed*)
- Exposes the triple store / SPARQL endpoint at public URL

Only data which are described as RDF triples can be harvested (mostly not the case)

- Harmonization of metagenomic datasets is illusory at the moment, checklists are not linked to the ontology terms => tedious manual work on the metadata side


https://data.emobon.embrc.eu/
«harvest only dataset info»
«harvest all emobon triples»
«custom harvest»
IDDAS
VRE
UDAL
SPARQL endpoint
custom App

Courtesy of VLIZ

Uniform Data Access Layer (UDAL)

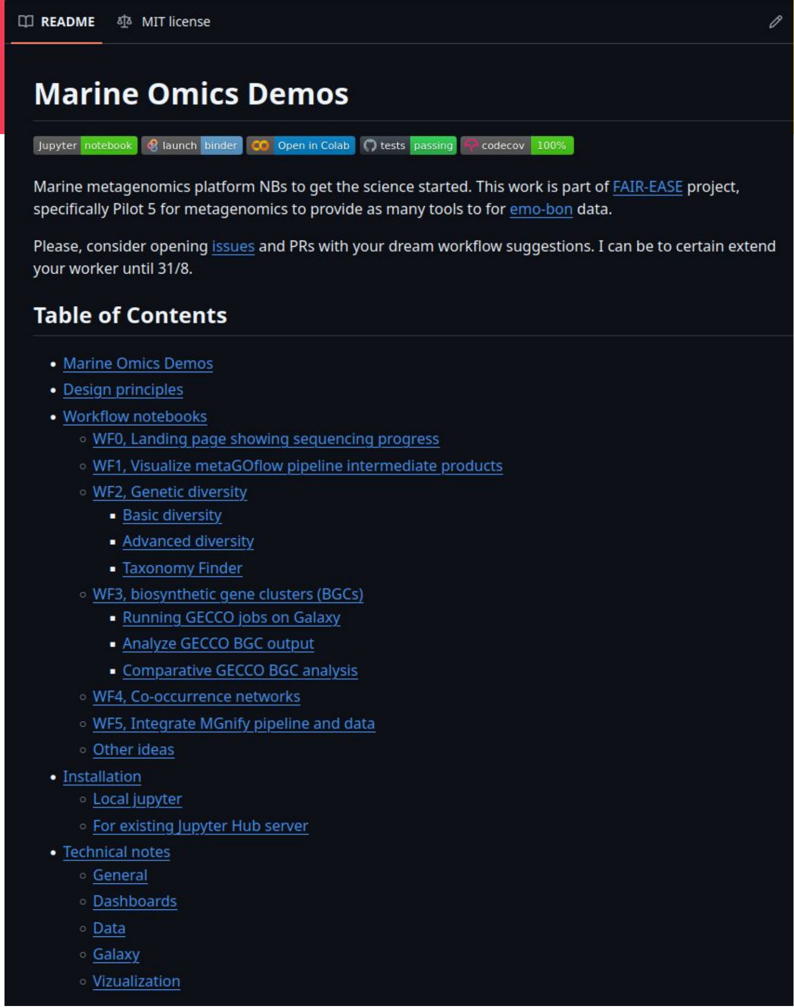
What are we solving: Data access

- replaces knowledge of SPARQL queries and data structure knowledge
- UDAL provides a agnostic interface for queries EMO-BON data across different sources and formats.
- Predefined set of queries, which get translated and delivered to SPARQL endpoint
- more explanation https://lab.fairease.eu/udal/
- what user receives is filtered dataframes, such as *pandas*

Data Analysis Kit

What are we solving: Exploratory data analysis

- Jupyter Notebook based dashboards
- https://github.com/emo-bon/momics-demos
- build on top of methods which should be applicable to other MGnify and MGO datasets
- https://github.com/emo-bon/marine-omics-methods
- Final deployment as VRE on blue-cloud 2026
- MIT licence, feel free to raise issues



Marine Omics Demos

Jupyter notebook | launch binder | Open in Colab | tests passing | codecov 100%

Marine metagenomics platform NBs to get the science started. This work is part of FAIR-EASE project, specifically Pilot 5 for metagenomics to provide as many tools to for emo-bon data.

Please, consider opening issues and PRs with your dream workflow suggestions. I can be to certain extend your worker until 31/8.

Table of Contents

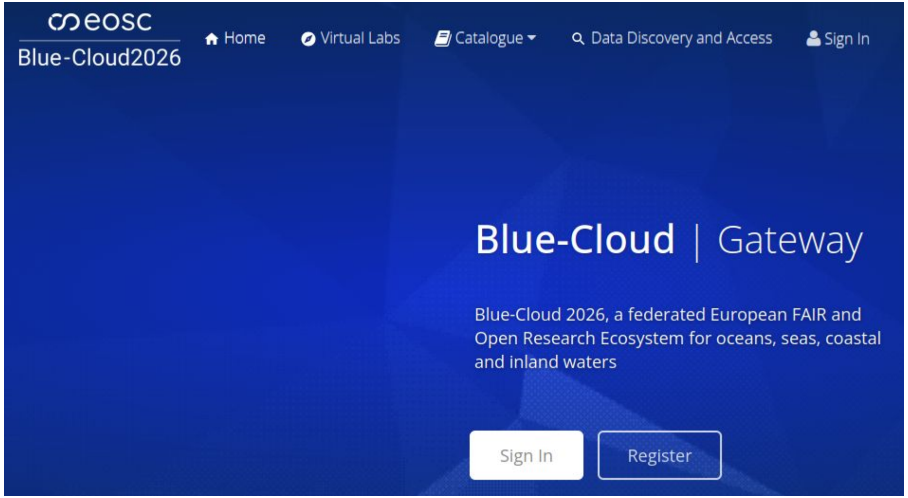
Ready for BioData.pt Management?

Financiado pela União Europeia NextGenerationEU

Blue cloud deployment

What are we solving: Integration of the whole

- Blue-Cloud was selected as one of the EOSC nodes, https://eosc.eu/building-the-eosc-federation/eosc-node-digital-twin-of-the-ocean/
- Many virtual labs already present
- Marine omics with the VRE will follow briefly.
- **IDDAS / DCAT** in blue-cloud as example of many existing data catalogues
- **Integrated Galaxy** instance to run workflows.



Blue-Cloud | Gateway

Blue-Cloud 2026, a federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters

What can I offer you NOW: Workarounds

RO-Crates

- Browsable in the RO-Crate viewer, not yet all 181 published samples (fastq files at ENA) from batch 1 and batch 2 (2021)

SPARQL

- For EMO BON, we compiled all the analysis results into singular tables (**parquet files**), plus combined **metadata table**.

Data analysis toolkit

- Installable locally / organizational jupyter Hub or limited user experience on public servers, mybinder.org and https://colab.google/
- No-code dashboards (*holoviz panel*) or interactive Jupyter notebooks using local UDAL implementation (https://github.com/fair-ease/py-udal-mgo)
  - Sequencing progress
  - Parsing and visualizing intermediate metaGOflow outputs (parsing RO-Crate metadata.json)
  - Alpha / beta diversities
  - Taxonomy finder
  - Biosynthetic Gene Clusters, enables submitting GECCO to Galaxy and visualizing results
  - Complex microbial networks based on taxonomic correlations
- **Harmonization**: Some NBs under development are trying to integrate results from MGnify + metadata from biosamples / ENA.
- Community inputs should pave the way for further development

Intermediate QA before the live demo

GLIM
BioData.pt

Ready for BioData.pt Management?

FCCN serviços digitais fct
fct Fundação para a Ciência e a Tecnologia
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU

Demonstration

- The slowest version, mybinder.org
- https://github.com/emo-bon/momics-demos
- 2GB of RAM, your laptop can certainly do better
- For local setup, https://github.com/emo-bon/momics-demos?tab=readme-ov-file#installation
- As on open-source developer, I am committed to improve the tools to community needs
- I am also compiling material for a comparative paper about several marine microbiome campaigns and EMO-BON, you are welcome to join forces.

GLIM

BioData.pt

Ready for BioData.pt Management?

More and all the questions?

dpalecek@ualg.pt
GH: palec87

BioData.pt
elixir PORTUGAL
FCCN serviços digitais fct
fct Fundação para a Ciência e a Tecnologia
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU