FAIR Data Management for Genomics Observatories

Katrina Exter (VLIZ), Cedric Decruw (VLIZ), Georgios Kotoulas (HCMR), Dimitra Mavraki (HCMR), Matthias Obst (UGot), Christina Pavloudi (HCMR), Marc Portier (VLIZ),Lennert Tyberghein (VLIZ)

What are we doing?

We are working on the management of the data from several marine Genomics Observatories. GOs collect physical samples, which are turned into digital data, which are analysed for the presence of species. The data collected from GOs are varied, and multiple lines of linked evidence can point to detected species.
Our challenge is to manage the flow of the data in such a way that the scientists doing the sampling, making the measurements, analysing the data, archiving and cataloguing the outputs, and using the data, can do so freely and easily, without breaking a sweat.

Why are we doing it?

The strength of GO data lies in the standardised and regular collection of physical samples, producing data that lead to species detections. Our GO data allow us to track the changes in marine populations with time and place, providing an important resource for studying the effect of climate change. However, GO data can be complex, with 100(0)s of sampling events spread over time and location, each with dozens of different types of measurements, and even with multiple lines of separate-but-linked evidence for the species detections which need to be analysed as a whole. Added to this, different GOs collect similar data but in slightly different ways.
We want to make it possible for any scientist to be able to find, use, understand, and combine the outputs from any GO, and to be able to combine the data with other long-term measurements made at the same sites.

We want to allow scientists to be creative in <u>what</u> they are doing, by freeing them from worrying about <u>how</u> they are doing it.

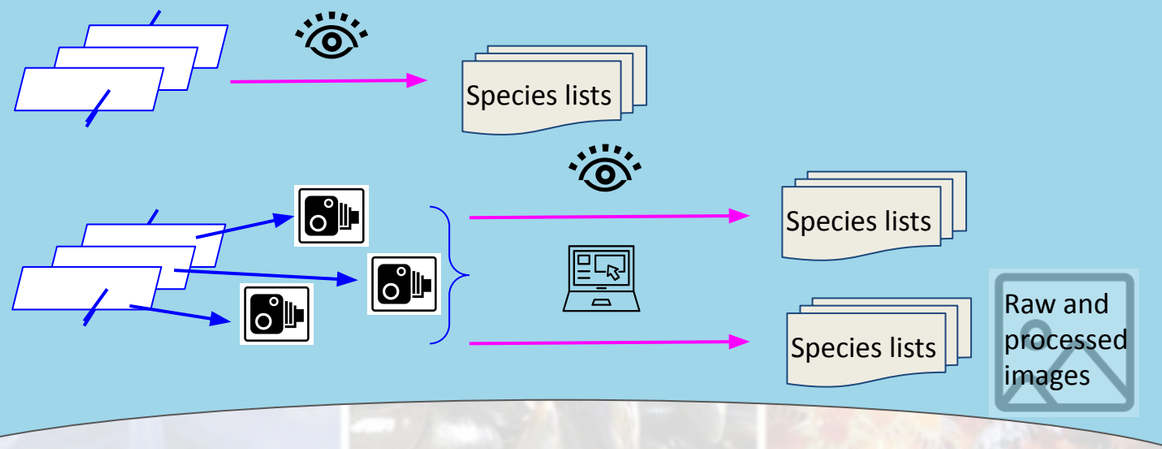
What is a Genomics Observatory (GO)?

GOs are ecosystems or sites that are subject to long-term monitoring of genomic biodiversity. This involves the regular, standardised collection and analysis of physical samples taken from the sites in the GO, which then produce observation, image, genomics, and physio-chemical data. These produce lists of species present at the GO site; as a whole the data are used to track changes in populations over time and location.

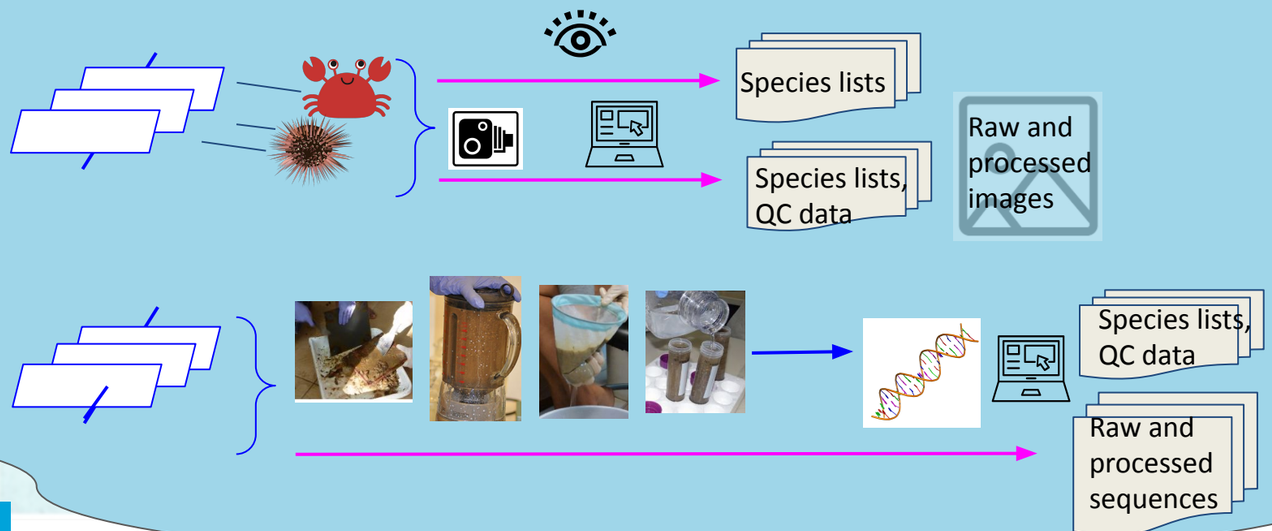
The OSD and ARMS-MBON GOs

Ocean Sampling Day and the *ARMS-MBON* project are two marine GOs that we are involved in. **Ocean sampling day** involves the regular (yearly, soon to be monthly) collection of water samples from dozens of marine sites scattered all over the world. The **ARMS-MBON** is a project of over 20 sites in Europe and the Ant(arctic). The Autonomous Reef Monitoring Structures (stacked settlement plates) are placed on the sea floor to be colonised by whatever lives nearby. Analysis of the ARMS and OSD data from DNA, images, and visual observations, and co-measured physio-chemical values, allows us to study the composition and functional potential of the local populations, *and* to track their changes over time and place.

leftmost ARMS units → sample method (visual observations [eye], photos [camera], or filtering and blending [on photos]) → analysis method (visual observation [eye] / image analysis[computer] / DNA analysis [computer]) → *rightmost* outputs (species lists, images, or DNA data)



An illustration of the four pathways ARMS: from samples to data



Our goals for: data life-cycle management

- Capturing data from the field: digital logsheets with FAIR-proof metadata
- Data archiving: easy, automatic storage of raw and (linked) processed data and all associated data
- (Meta)data cataloguing: automatic creation of rich metadata records
- Provenance (meta)data management: starting from the field

Our goals for: data processing management

- Versioning and timestamping
- Applying workflows for data analysis, allowing for m2m interactions
- Applying semantics and using controlled vocabularies
- Provenance: ensuring that links between raw and processed data, products, and results is done efficiently

Our goals for: engagement management

- Capturing metrics, comments, and derived results and corrections
- Writing short and sweet HowTos, cheat-sheets, to overcome the human resistance to reading documentation
- Easy to fill forms, with error-proofing!

Our goals for: creating rich data products and data explorers

- Creating DwC-A products to publish, but also exploring other linked-data formats
- Data explorer: filter and select to find data, explore within the datasets to identify more specifically what there is and if you want it

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

EMBRC

VLIZ
Vlaams Instituut voor de Zee vzw
Flanders Marine Institute