FAIR Data Management for Genomics Observatories

Katrina Exter, VLIZ (BE), katrina.exter@vliz.be
Cedric Decruw, VLIZ (BE), cedric.decruw@vliz.be
Georgios Kotoulas, HCMR (GR), kotoulas@hcmr.gr
Dimitra Mavraki, HCMR (GR), dmavraki@hcmr.gr
Matthias Obst, UGot (SE), matthias.obst@marine.gu.se
Christina Pavloudi, HCMR (GR), cpavloud@hcmr.gr
Marc Portier, VLIZ (BE), marc.portier@vliz.be
Lennert Tyberghein, VLIZ (BE), lennert.tyberghein@vliz.be

Genomics Observatories (GOs) are an increasingly important resource to study the effect of climate change on marine populations. The data gathered by GOs allow one to map and track how marine populations change and how those changes relate to the local and global conditions. Such data may be used to calculate Essential Biodiversity Variables (EBVs) and can provide important information for predictive modelling of marine biodiversity.

To take full advantage of data from GOs – whoever produces the data – it is necessary that their data are FAIR: the data are findable in community catalogues, they can be accessed by human and machine, they use community standards in the data formats and the metadata vocabularies, provenance is fully documented; and ideally – but not necessarily – the data are also open access. Taking into account the complexity of GO data – with (a)biotic, genomic, geographic, and etc. parameters that need to be linked in a humanly-understandable but machine-interoperable way – it is clear that solid thinking and planning about the management of the data is essential. Done well, this allows scientists to be creative in what they are doing, by freeing them from the how of what they are doing.

VLIZ is involved in a number of GO projects, and here we explain the steps we are taking in managing the data to satisfy current and future scientific needs.

Two of these GOs – ARMS (Europe) and Ocean Sampling Day – operate (partially) under the umbrella of ASSEMBLE Plus, and we are working with EMBRC to ensure the long-term sustainability of these marine GOs.

- European ARMS programme (Matthias Obst, UGot): long-term monitoring and biodiversity assessment of invasive and indigenous hard-bottom communities. Running as collaboration of dozens of institutes, this network of *A*utonomous *R*eef *M*onitoring *S*tructures (ARMS) are deployed in the vicinity of marine stations and LTER sites in Europe and Ant/arctica for a period of 3-24 months at a time. Visual, photographic, and genetic assessments are made of the communities that settled on the ARMS units: these data will be used to track the species populations, in particular sensitive and invasive species, to map migrations, and to identify EBVs for hard-bottom fauna. ARMS units have been deployed each year since 2018 and currently have a time-series of 3 years of data.
- Ocean Sampling Day (Georgios Kotoulas, HCMR): a simultaneous sampling of the worlds oceans on the summer solstice of each year. OSD began in 2014, and continued under ASSEMBLE Plus from 2018 onwards. Following standardised sampling and sequencing protocols, these data allow for an assessment of the species populations. Packaged together with the collected (a)biotic parameters, standardised datasets will be produced that can be compared over space and time. An extension to a monthly sampling campaign, sponsored by EMBRC, will also soon begin.

The data from these two GOs consist of

- Sequences: metabarcoding and shotgun metagenomics → analysed to produce information about the diversity of species and populations, and about the metabolic diversity of prokaryotic communities
- Images and visual assessments of the ARMS plates → analysed to produce species and biomasses/abundances
- Abiotic and biotic parameters → analysis of environmental parameters

To take full advantage of these projects, it is a must that the ARMS and OSD datasets from each station and each year can be compared, that the data can always be linked back to each sampling event and their unique samples, and that all data can be loaded into statistical codes and any variety of data-analysis workflows and virtual research environments (VREs). *But we also want that the ARMS and OSD datasets can be combined with each other – even with other data from other projects* – and this should be possible for any scientist to do, not only for those intimately involved in the projects. *In short, the data need to be FAIR*.

To this end, we are tackling the following aspects:

Data life-cycle management

- Capturing the data from the field: helping the scientists create their digital logsheets with the necessary data and metadata, and using standard vocabularies from the very beginning
- Data archiving: ensuring easy, automatic, and permanent storage of raw and processed data, permits, protocols, etc.
- (Meta)data cataloguing: automatic creation of rich metadata records for the OSD and ARMS data, with an organisation that follows the life-cycle of these projects
- Provenance: ensuring provenance is fully included as (meta)data

Data processing management

- Versioning and timestamping
- Applying workflows for data analysis, allowing for machine2machine interactions with the archives and catalogues
- Applying semantics and using controlled vocabularies
- Provenance: ensuring that the links between all and any raw or processed data, data products, scientific results, and their provenance, is done with as little human work

Engagement Management

- Capturing metrics, comments and derived results and corrections
- Writing short and sweet HowTos and cheat-sheets, to overcome the oh-too human resistance to reading documentation; and creating templates with pre-selected intelligent suggestions, to minimise the resistance to filling in forms!

Creating rich data products and data explorers

- Creating Darwin Core (OBIS Event) data products to hold all information: linking the identified species to the data from which they were identified back to the samples from which the data were obtained
- Exploring these DwC files with a user-friendly explorer: to allow the entirety of the OSD and ARMS data to be visualised, inspected, selected, grouped, cross-matched, and explored.