BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e58938
doi: 10.3897/biss.4.58938

OPEN ACCESS

Conference Abstract

Genomics Observatory Use-Case: The challenge to standardise image and sequence data to Darwin Core format

Katrina Exter[‡], Cedric Decruw[‡], Marc Portier[‡], Vasilis Gerovasileiou[§], Christina Pavloudi[§], Matthias Obst[|]

‡ Flanders Marine Institute, Oostende, Belgium
§ Hellenic Centre for Marine Research, Heraklion, Crete, Greece
| Göteborg University, Gothenburg, Sweden

Corresponding author: Katrina Exter (katrina.exter@vliz.be)

Received: 24 Sep 2020 | Published: 09 Oct 2020

Citation: Exter K, Decruw C, Portier M, Gerovasileiou V, Pavloudi C, Obst M (2020) Genomics Observatory Use-Case: The challenge to standardise image and sequence data to Darwin Core format. Biodiversity Information Science and Standards 4: e58938. https://doi.org/10.3897/biss.4.58938

Abstract

Genomic 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 with time and location, 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.

GOs are sites that are subject to long-term scientific research, including (but not limited to) the sustained study of genomic biodiversity from single-celled microbes to multicellular organisms. We are involved in a number of GO projects, including:

- Ocean Sampling Day (OSD): yearly, standardised water sampling from selected sites located all over the world. The data collected consist of (a)biotic parameters, DNA sequences extracted from the water samples, and species occurrences derived from analysis of the sequences.

- Marine Biodiversity Observation Network of Autonomous Reef Monitoring Structures (ARMS-MBON): yearly placement of sets of stacked plates along the harbours and coasts of Europe, deployed in place for months to allow species to settle on them. The data collected consist of images of the communities that settled on the plates, DNA sequences extracted from scrapings of the plates and the surrounding water, visual observations, and species occurrences derived from analysis of the sequences and the images.

The data collected from these GOs are not particularly complicated: sampling and sequencing protocols are well established and the images are taken with standard cameras. However, complexity arises when

1. measurements need to be linked between multiple samples extracted from *each* event for *each* location and date, to *all* the data from *all* sampling events, and
2. the results—species occurrences and abundances—obtained from the images *and* the sequences for each sample, must be linked together within that sampling event and between all sampling events of the GO project.

We also want to be able to compare data between different GO projects, and to be able to incorporate measurements from nearby monitoring stations, as this will allow for an enhanced analysis of the evolution of marine benthic populations in light of climate change.

Our aim is to adopt the Darwin Core Archive (DwC-A) OBIS-ENV-DATA format for the data from ARMS-MBON and OSD, including the linkages to the images and sequences. ARMS-MBON presents a challenge because of the wide range of data collected. Our use-case for the data format we require contains the following elements:

- The ARMS-MBON data that are collected are the sequences obtained from the samples and the images taken of the communities on each ARMS plate. Species occurrences and added (a)biotic parameters only come later. However, we would like to adopt the DwC-A format already being widely used from the very beginning of our data management, with the species occurrences being added to the same DwC-A files when they are determined.
- The ARMS plate images will go through a few stages of processing: the raw images taken by the field scientists, and later annotated images created by the image-analysis software. We would like to link the raw and processed images to each other within the DwC-A files, so the user can obtain the processed image a species was identified in, and the raw image for its re-use. This same requirement applies to the sequence data.
- The ARMS plate images from each sampling event number many dozens. These are archived as ZIP files in the Marine Data Archive, and the links to them are made public via the metadata record in a catalogue. We would like to be able to add links in the DwC-A files to the individual images in the ZIP file that each species occurrence came from, and from which region(s) on the images they were found; so individual images can be referenced but the space saving advantage of ZIP can be used.

- In addition, we would like to explore the option of adding the images *within* the DwC-A (ZIP) file itself, rather than referencing them as URLs.
- For each sampling event we collect sequences and images, and species (occurrences) will be obtained from analysis of both types of data. It is necessary that these multiple and overlapping results can be linked to each other in a clear and consistent way. This means indicating clearly where sequences and images are of the same or of different material from any plate.

Clearly, a high degree of data standardisation is necessary to enable harmonisation of the data, to accommodate the multiple streams of linked data values, which will be added by multiple creators and users of the data.

Keywords

genomics observatories, Darwin Core Archive, imaging data, DNA sequences, species occurrence

Presenting author

Katrina Exter

Presented at

TDWG 2020

Funding program

ASSEMBLE Plus: OSD and ARMS receive funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730984.

EMBRC: The OSD and ARMS projects are supported by the European Marine Biology Resource Centre

EMODNet Biology: This work has been financially supported by the EC DG-MARE (EMODnet Observation and Data network - Lot5 - Biology: EASME/EMFF/2016/1.3.1.2/Lot5/SI2.750022).