Contemporary Data Management for Biodiversity Observation Networks leading to Linked Open Data Publishing through Distributed Techniques applying RO-Crate and GitHub Actions

Marc Portier‡, Cedric Decruw‡, Katrina Exter‡, Rory Meyer‡, Lennert Tyberghein‡, Laurian Van Maldeghem‡

‡ VLIZ vzw (Flanders Marine Institute), Ostend, Belgium

Corresponding author: Marc Portier (marc.portier@vliz.be)

Abstract

Biodiversity Observation Networks (BONs) are important sources of information about the health and wealth of biodiversity in our world. By observing the presence or absence of species in different environments and by coupling with measurements of environmental parameters, they contribute important information to our knowledge of the natural world and to our ability to model and predict the effects of climate change. Knowledge gained from BONs can also be used to assist wider society (political, social, and economic bodies) in making biodiversity-sensitive decisions.

For the data outputs from BONs to be useful to a wide audience, management of those data is crucial: the data should be published openly and they *must* be Findable, Accessible, Interoperable, and Reusable (FAIR). It is important that the data can be found and understood by anyone who could use them. For BON data to lead to new science, it is necessary that the data can be accessed programmatically, that full provenance is provided, and that data from different sources are interoperable. Semantic and technical interoperability should be coupled with user-friendly data archiving, managing, discovery, and User Interfaces (UIs) for those creating and for those using the data. The importance of lowering the hurdle to creating and sharing data in a FAIR way is not to be underestimated; the hardest part of the FAIR journey is often its uptake rather than its technology.

The importance of eDNA (environmental DNA) to BONs, highlights the particular challenge to the management of BON data because of the complexity of the practical workflow (many parties involved doing the actual sampling, samples being shipped as well as biobanked, highly technical procedures involved at various stages), the complexity of the analysis of the DNA to measure “occurrences”, and the rapid evolution of the field of biotechnology.

Version management to accommodate updates to biotechnology pipelines and reference databases, the use of IDs to allow data to be linked to evolving *knowledge* rather than to static information, semantic annotation aimed at multiple audiences (omics-expert and less-expert), and linking multiple-distributed datasets (omics, image, environmental, etc.) are all crucial aspects to FAIR management of omics data. As data standards evolve much more slowly than the scientific possibilities, it is also important to archive, annotate, and link all the data in a flexible way so it can be exported into today's *and* tomorrow's data formats. All of the above expectations are added as extra weight onto the field scientists and lab workers dealing with the actual samples and producing their digital outcomes.

At the Vlaams Instituut voor de Zee (VLIZ) we are managing data from several BONs, and for this we are developing a data management approach that is based on Research Object Crate (RO-Crate) data packaging including the following elements: a GitHub entry point for holding the RO-Crates, uploaded data (co-located files or links to data held or published elsewhere), a straight-forward GitHub-Action workflow to publish and uplift the contents as linked open data, and UIs for data upload, search, and select. BON contributors and data managers will be able to upload the BON data (e.g., sampling log sheets, ENA (European Nucleotide Archive) accession codes, SOPs (Standard Operating Procedures), etc.), interfacing directly with GitHub or through our assisting tool to upload and create the RO-Crate descriptions, and to semantically describe their data. Scientists who analyse data will also be able to upload their workflow outputs and provenance metadata descriptions and link those to the BON data that their analysis is based on. By collecting the data and metadata in one place, the BON outputs can be rearranged into whatever published formats are required.

The advantages of this approach are multiple:

- It relocates the authority on "what can and should be said about the data" back to the authors, who are not then limited by the data formats of the publishers. The application profile and required fields they themselves agree upon allow for expressing all the nuance and targeted information that they want, and thus create an overall stronger affinity between the produced content and its maintainers.

- Embracing the semantic web techniques marries a high-level uniform processing and indexing with a diversity of added detail that can be retrieved by drilling down. Similarly, clear cross-references to managed terms in managed vocabularies allows further serendipitous connections to apparently unconnected domains.

- Providing digital assistance allows automated provenance tracking, easy linking to agreed managed vocabularies, as well as hiding cumbersome low-level technical details (such as Git operations, format specifications, etc.). By doing so as early as possible in the processing of the information, it avoids lost memory details, late quality checking, and lengthy round trips that all add to the cost of publishing, interpreting and reusing the produced data.

- Finally, the available open semantics in the datasets themselves allow for advanced additional post-processing that is not limited to that provided by centralised aggregation services, but also allows for providing fragmentations, indexes and navigational indicators that support a more scalable, distributed and federated real-time pathway to selecting and analysing data in the context of specific research questions.

Keywords

FAIR data

Presenting author

Marc Portier

Presented at

TDWG 2022

Conflicts of interest