

OBIS 2.0: Towards real-time integration, quality control, and analysis of rich marine data streams

Appeltans Ward¹, Provoost Pieter¹, Bosch Samuel¹, Bristol Sky² and Klein Eduardo³

¹ OBIS, Intergovernmental Oceanographic Commission (IOC) of UNESCO, IOC Project Office for IODE, Wandelaarkaai 7/61, 8400 Oostende, Belgium

E-mail: w.appeltans@unesco.org

² OBIS-USA, United States Geological Survey, W 6th Ave Kipling St Lakewood, CO 80225, United States

³ Caribbean OBIS, Universidad Simon Bolivar, Depto Estudios Ambientales AP 89000, 1080 Caracas, Venezuela

Millions of marine species observations as well as measurements of abundance and biomass can be accessed through the Ocean Biogeographic Information System (OBIS). In recent years, OBIS developed a RESTful Application Programming Interface (API) to help drive a new portal interface and an R package for analytical use, which is proven to be attractive to users and now drives multiple third-party applications. However, the system for harvesting new data was experiencing major delays due to aging and out of date processes and software infrastructure. The process of quarterly data loads through a process that involves a fair bit of manual labor in addition to long processing times to rebuild database indexes and other aspects of the system was problematic in getting new data online and meeting user expectations - from both data providers and data consumers.

Here we present our new OBIS infrastructure and technology stack which aims to achieve the following high-level objectives:

1. Near real-time data integration from OBIS Nodes - as soon as OBIS Nodes advertise the availability of new or updated data, the OBIS system begins automated processing to ingest data and completes integration routines in a timely manner.
2. The ability to scale the system to hundreds or thousands of new datasets and millions of new records while providing timely results for queries, data access (download and streaming), and products (analytics and indexes).
3. The ability to integrate environmental variables in addition to biological observations and to fully leverage this new type of data in queries, data access, and products.
4. Improve on the real-time analytics capability of OBIS to directly inform biodiversity indices, essential ocean variables and biodiversity indicators, starting at the API level to support science end users (e.g., via R/Python packages) and custom apps/portals along with a flexible and ever evolving set of reports through the OBIS portal.
5. Improve the full visibility of the OBIS Network by persistently linking OBIS Nodes, data provider institutions, and individual contributors (data providers, authors, etc.) together such that each entity is cited and their full contribution to OBIS easily accessible.

Future plans include an online data entry and editing tool including the capability to manage both open, public data as well as data that are under limited embargo due to research project constraints and the ability to provide and share annotations on the quality of the published data.

Keywords: OBIS; biodiversity data; software infrastructure; quality control; FAIR principles