OBIS Products Catalogue: Development Plan

The Intersessional Working Group on Data Products highlighted the need for a catalogue in which data products developed with OBIS data by the Secretariat and the nodes can be easily found and accessed (see annex). This initiative will also help ensure proper acknowledgement and reporting of the work carried out by our community. The subsequently established Products Coordination Group likewise endorsed the catalogue as a top development priority.

Here, we present a development plan for the catalogue, outlining the required steps and identifying several points that still need to be discussed with the coordination group.

Timeline

- October 2024 - October 2025: Exploration and development of CKAN
- June 2025: Expansion of CKAN requirements to include JSON-LD for provenance tracking
- October 2025: Initial ingestions of test datasets
- SG13 (October 2025): Beta testing of platform
- November 2025 - December 2025: improvements based on experience during SG13 testing.
- December 2025: Development of User guide
- January, 2026: Soft Release - submission open to OBIS nodes
- February, 2026: Release 1.0: submission open to OBIS community

Publicity

- How are we going to make known that we have a catalogue?
    - News on OBIS website, social media
    - Possible promotion on IODE and IOC social media and websites
    - Circulate to OBIS community through newsletter
- How do we promote the usage of this resource?
    - Create articles for the OBIS nodes newsletter
    - Create section on OBIS manual or dedicated training material
    - Further development of documentation (user guide; see below)
    - Directly invite researchers/institutions with recognized/widely used products to publicize it also on the catalogue, bringing some attention to the platform

Documentation

- User Guide
- Feedback on user guide: https://github.com/iobis/obis-products-catalog

Requirements

Metadata

- DataCite / Zenodo
- CKAN: https://github.com/ckan/ckanext-scheming
- Schema.org, ODIS Flavor: https://book.odis.org/thematics/index.html

Quality

- Minimum quality requirements: to be discussed on next PCG meeting
- Types of products that will be accepted: to be discussed on next PCG meeting

Products updates

- Main mechanism is DOI registry, with periodic reharvest
    - Frequency of reharvest is still to be decided given technical requirements

Submission Mechanisms

- DOI ingestion
- Manual submission

In both cases it is still to be discussed if this will be administered by the secretariat (helpdesk) or by nodes.

Review Process

Two options are available, which will be further discussed in the PCG:

- No review
- Direct submission with review or review prior to CKAN submission

User Management

- SSO eventually, including Google/ORCID
- Possibility of link with OceanExpert

Linkage to other catalogs (e.g. ODIS)

- JSON-LD schema - internet (e.g. Google)

- ODIS

Sustainability Plan

- Currently there is no short-, mid- or long-term plan for the financial maintenance of the catalogue. While the platform, per se, does not demand significant technical resources, the administration, promotion and updates will certainly require personnel.

Annex I

Concept note for a products development and sharing infrastructure for OBIS, developed by the Intersessional Working Group on Data Products (predecessor of the PCG).

OBIS data products

Concept note

Outcome of the IWG on OBIS data products

Contributors (in alphabetical order)

Elizabeth Lawrence, Hanieh Saeedi, John Nicholls, Johnny Konjarla, Jon Pye, Katherine Tattersall, Lisa Benedetti, Pieter Provoost, Sachit Rajbhandari, Salvador Fernandez, Silas Principe, Stephen Formel, Takashi Hosono, Ward Appeltans, Yi-Ming Gan

March 2024

Background

OBIS developed the **OBIS2030**, an UN Ocean Decade endorsed project, to provide a biodiversity data hub made up of standardised, quality controlled and managed data to support the Ocean Decade objectives. This will help researchers, practitioners and decision makers to protect and restore marine ecosystems and protect life in the ocean.

One integral part of the OBIS2030 targets is to **create and publish information products, at global, regional and national scale to feed directly into reporting and assessment processes.** This should occur by providing a platform to (jointly) develop and share reliable biodiversity indicators and information products that describe changes in marine ecosystems.

To start with the implementation of those solutions, the SG-OBIS decided to establish an open-ended intersessional working group on OBIS-based marine biodiversity indicators and information products (IWG-OBIS-PRODUCTS).

The IWG-OBIS-Products is a collaborative and interdisciplinary group that is driven by the importance of creating indicators and information products that are **scientifically sound**, practical, and relevant to decision-makers in government, industry, and civil society. Our main objectives are to:

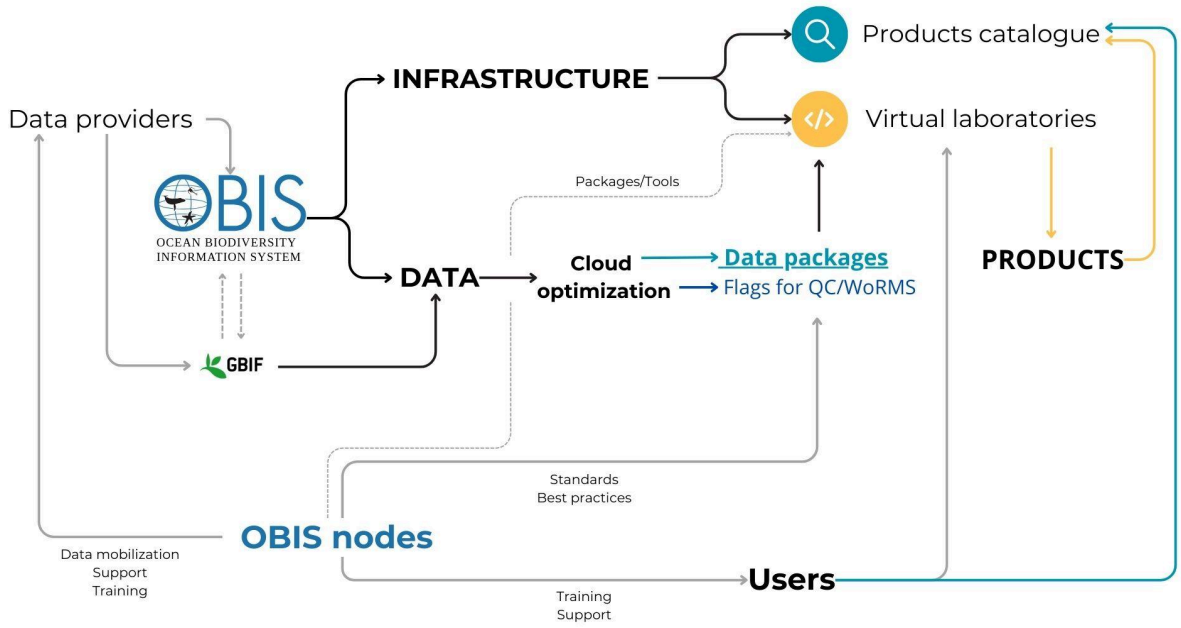
- Review existing OBIS information products
- Identify products needs (at local, regional or broader scales) in order to create a plan for future products implementation by contacting nodes and stakeholders, including the community of users
- Perform a systematic analysis of the data in OBIS, identifying geographic and taxonomic gaps, which can help identify indicator species and ecosystems or strategies for species distribution modelling
- Contact ecological synthesis centres/groups to promote an exchange of knowledge and gather suggestions/advice on possible products that could be derived from OBIS
- Propose a process for frequent expert validation of information products by consulting with local scientific experts and end-users (including local communities and indigenous people)
- Discuss the relevance of supporting community generated OBIS products and establish a potential framework for receiving and portraying those products.

Part of those objectives were already completed and generated this concept note.

Proposal

Expanding the toolbox of OBIS users

When the IWG was established, our initial proposal was to have a catalogue to showcase products derived from OBIS data. However, as discussions took place, it became clear that there was an opportunity to create a whole set of tools that would enable users to work and take the maximum advantage of OBIS data more efficiently. The group then proposed to create a new structure based on **infrastructure** and **data** components. This infrastructure will enable the production and publishing of **data products based or partially based on OBIS**.



Infrastructure

The infrastructure will include (1) a virtual laboratory (phase 1) and (2) a products catalogue (phase 2). The virtual laboratory is the main tool we are going to provide and is a **JupyterHub** with Python and R kernels, including an RStudio interface (like an RStudio server). This will provide users the opportunity to run their analysis online, taking advantage of an environment containing tools and cloud optimised data (for example, the full export of OBIS and data packages, see on “Data” section). Users would have available an environment containing the main packages used for biodiversity analysis and one

dedicated package to speed up pre-processing of OBIS data for specific purposes. Also, model scripts would be available, with the nodes contributing in both cases.

Since data and analytical tools reside on the same machine (the server), analyses are faster, and the user does not need to download any additional data or tools. This type of infrastructure is mature and is being successfully used by other organisations (e.g. Copernicus/WeKEO: https://www.wekeo.eu/; Digital Earth Australia: https://www.dea.ga.gov.au/developers/sandbox). This is also aligned with the Digital Twins of the Ocean idea, which aims to speed up the production of data products from ocean observation.

We consider data products any type of analysis (description, data visualisation, etc.) that synthesises and generates new information from data hosted on OBIS and other sources. Some examples are the contribution of OBIS to the State of the Ocean Report (see pages 26 and 27 of the report); species distribution maps such as AquaMaps (https://www.aquamaps.org); and the Marine Biodiversity Observation Network (MBON) early alert dashboards (https://marinebon.org/data-products). Those are just illustrative examples, as the potential uses are multiple. One particular type of product that the IWG perceives as a priority would be data visualisation, as those are constantly requested by the community and can be used by multiple stakeholders.

In a second moment we would also implement a products catalogue, to showcase those products being produced through the virtual environment, but also elsewhere. This was a particular demand received from some stakeholders, as they want to contribute with OBIS products, but have their own systems/websites. Also, this would enable users with less familiarity with programming languages and JupyterHub to access and use the products. Our first suggested solution was **CKAN**, but we will also test the **GeoNetwork** platform. All products on the catalogue will be accompanied by extensive metadata, in a way that it could be easily integrated with ODIS.

JupyterHub → JupyterHub is a multi-user server environment designed to facilitate collaborative and interactive computing in an academic or research setting. It serves as a platform for deploying Jupyter Notebooks, which are open-source, interactive web applications that allow users to create and share documents combining live code, equations, visualisations, and narrative text. JupyterHub, as a centralised system, enables multiple users to access and utilise a shared computational infrastructure concurrently. This centralised approach enhances efficiency, scalability, and resource management in academic environments where collaborative data analysis, scientific computation, and research are paramount.

CKAN → The Comprehensive Knowledge Archive Network (CKAN) is a robust and extensible open-source data management system widely employed in academic and research contexts. Functioning as a data portal platform, CKAN facilitates the storage, management, and dissemination of diverse datasets. Its architecture is designed to support the systematic organisation of data resources, ensuring metadata quality, version control, and accessibility. CKAN serves as a centralised repository, empowering institutions, researchers, and policymakers to efficiently share, discover, and access datasets. Its modular and customizable nature, coupled with a rich ecosystem of extensions, renders CKAN a versatile solution for institutions seeking a standardised, transparent, and collaborative approach to data management and dissemination within the scholarly domain.

Data

On the data side, the proposal is to provide users of the virtual environments with **cloud optimised data**, providing a seamless and faster analysis. Also, we would provide users with data packages - pre-filtered data targeting specific uses. Example, a package of data including only long-term monitoring, or all eDNA data. Of course, for some studies the data package will still need some treatment, but hopefully it would maximise the effort of researchers by skipping some steps.

As a starting point, we would leverage recent developments created for projects and partnerships that OBIS is involved in, such as the State of the Ocean Report (StOR, in partnership with ProtectedSeas; https://github.com/iobis/protectedseas-statistics), the eDNA expeditions (https://www.unesco.org/en/edna-expeditions) and the MPA Europe (https://mpa-europe.eu/). The StOR, for example, already applied one of the suggestions of the IWG team - to have the full export of OBIS indexed by the H3 system (a geospatial indexing system developed by Uber Technologies; https://h3geo.org).

It is well known that OBIS data presents a good quality, thanks to the efforts of OBIS nodes, and is already quality controlled in multiple ways. It was suggested that, when creating those data packages, we have the possibility to add additional QC steps focused on the data purpose (for example, on a long-term monitoring tag a year with a very different number of samples). Another suggestion is to make available scripts to perform this type of QC in standardised ways, saving scripting time for users, while sharing experiences among researchers.

OBIS nodes are key for the data component in two aspects: they are the point of contact with the local community, understanding data needs, and they provide best practices guidelines for QC of data. The idea is that OBIS nodes would be able to implement

routines (through scripts) to generate specific data packages, and those could be run in the OBIS server.

One of the concerns, as outlined in the 'critical points', is that for this data to be truly useful for the user, it must be consistently updated. This necessitates the creation of pipelines for automatic data updates in a timely manner. Naturally, all data should be accompanied by a versioning system to keep users informed of changes, and the generation codes should be openly available, providing users with the option to generate the data package independently.

Cloud optimised data → Cloud optimised data refers to a paradigm in data storage and organisation tailored for efficient and scalable utilisation within cloud computing environments. This approach involves structuring and formatting datasets in a manner that aligns with the distributed and parallel processing capabilities inherent to cloud platforms. Characterised by columnar storage, partitioning, and utilising formats conducive to parallel data processing, such as Apache Parquet, Zarr and COGeo, cloud-optimised data aims to enhance query performance, reduce data transfer costs, and facilitate seamless integration with cloud-based analytical tools.

Examples of products needs

The proposed infrastructure would enable the OBIS community, including the secretariat, to generate products that are constantly required by stakeholders. We identified some of those needs (with contributions of the community):

- Spread and distribution of non-native species over time
- Distribution of records by depth and bottom depth
- Animation of OBIS records through time and depth
- eDNA data dashboard
- Diversity indicators (with possible corrections)
- Harvesting marine datasets from GBIF

The IWG group agreed that an stakeholders assessment would be important as a next step, to identify the priority products, and also data and tools needs.

Critical points

The IWG also recognizes some challenges in the implementation of the proposed plan. Those are by no means impeditive, but will require a thorough consideration by the team.

- Frequency of update of data - to be meaningful, the data packages (or any full data export) need to have a reasonable update frequency. Of course, the frequency may vary depending on the data purpose. This will need automated pipelines.
- Infrastructure costs - while we plan to start/test the infrastructure with our available computational resources, in the mid/long-term it is expected that new resources will be needed. In this sense, there will be costs involved. Possible ways of tackling this problem are approaching partners to share infrastructure (e.g. LifeWatch) or finding specific funds for this project.
- Training needs - it is anticipated that many users may not be familiar with JupyterHub. Also, some of the OBIS users do not have knowledge of programming languages like Python and R. In that sense, it will be essential to provide training to ensure the progressive adoption of JupyterHub as a tool by the OBIS community.

Implementation plan

Step 1 - Test of infrastructure

- Implementation of a JupyterHub instance [done]
- Implementation of a CKAN or GeoNetwork instance
- Test of data packages
- Test of “recipe scripts”

Step 2 - Prepare tools and data packages (with OBIS nodes)

- Prepare an R package with main tools (specifically tailored for our infrastructure)
- Collect sample scripts that may be of interest
- Prepare a first set of data packages

Step 3 - Open to test with community

- Collect users perceptions
- Evaluate platform functioning and make adjustments

Step 4 - Open to use and implement the CKAN for showcasing products