OBIS Grand Unified Model Project Team Report

Distributed to SG-OBIS via email on September 26, 2023

Introduction

Expanding the data model (previously known as the Grand Unified Model) is an effort funded by the Global Biodiversity Information Facility (GBIF) to explore new ways of modeling the complexity inherent in biological data. The Ocean Biodiversity Information System (OBIS) community encounters data that is difficult to capture well using Darwin Core and developed the Extended Measurement or Fact extension (EMoF) to help address this issue. However, even with that extension there are still obstacles to being able to seamlessly model the data. Given the goal of the new model, the OBIS community was interested in exploring its capabilities and the OBIS Grand Unified Model Project Team (OBIS GUMPT) was formed at the May 2022 OBIS Steering Group meeting. Abby Benson (OBIS-USA) raised the issue of the new data model at that meeting and volunteered to lead the project team. Yi-Ming Gan (AntOBIS) volunteered to co-chair the project team. The purpose of the project team was for the OBIS secretariat and nodes to explore the new model via use cases that were most pertinent to the marine community and assess its capabilities.

Project Team Rationale

GBIF is currently designing a new conceptual data model and associated simpler data publishing models capable of supporting families of similar use cases (e.g., environmental DNA, camera traps, biotic interactions, biotic inventories, etc.). As noted previously, OBIS has provided textual content to the use cases such as the environmental and community measurements use case. The new model and data publishing subsets represent an opportunity for OBIS to provide direction and guidance into how the models can best represent OBIS community data and also an opportunity for OBIS to prepare for this new direction.

Project Team Initial Plan

The project team explored early adoption and testing of the new data model to assess how well it would work for OBIS community data, noting and sharing back to the data model team any problems encountered, suggestions for improvements, and feasibility of uptake.

The team identified the following tasks:

1. Select between one to five use cases that most clearly align with OBIS community data and that the project team has datasets ready to test for the use case.
    a. Identify datasets not covered yet by any of the use cases and feedback to GBIF.
2. Maine Inshore Trawl Survey dataset example
3. BRUV dataset to CamTrapDP
4. Apply the data model to the selected datasets.
5. Document issues, suggestions, and feasibility for each use case.

6. Explore the feasibility of using frictionless data packages instead of Darwin Core Archives.
7. Assess impact to OBIS data system including amount of work necessary, funding required, sources for funding if required, and recommendation on adoption.
8. Report our findings to SG-OBIS, TDWG, GBIF Global Nodes and the GBIF data model project team.

Membership

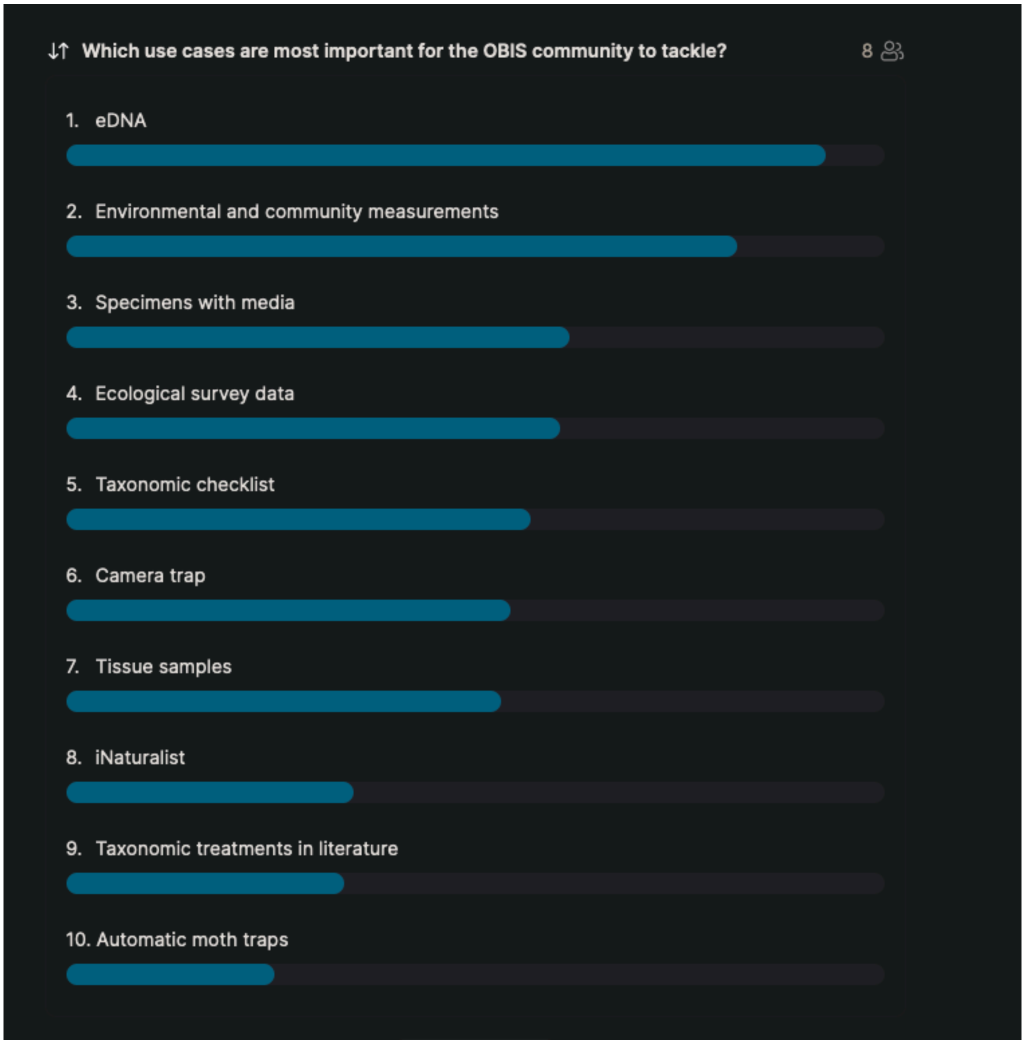
Abby Benson (Chair, OBIS-USA), Yi-Ming Gan (co-chair, AntOBIS), Pieter Provoost (OBIS Sec), Maria Cornthwaite (OBIS Canada), Ward Appeltans (OBIS sec), John Nicholls (OBIS-OPI), Saara Suominen (OBIS Sec), Martha Vides (OBIS- Co), Elizabeth Lawrence (OBIS Sec), Serita van der Wal (OBIS Sec), Katherine Tattersall (OBIS-AU), Sachit Rajbhandari (OBIS-AU), Ruben Perez Perez (EurOBIS), Kevin Paxman (OBIS-UK), Tim Robertson (GBIF Sec), John Wieczorek (GBIF contractor, Darwin Core lead)

Meetings

The project team met a total of eight times from June 2022 until October 2023. Meetings were generally held twice per day for the project team chair to facilitate global participation in the project team.

OBIS Identified Key Use Cases

Initially the project team used a Google Jamboard to identify datasets that project team members could think of related to each of the thirteen use cases and the project team determined that only ten of the use cases would have relevant data from the OBIS community. Then the project team evaluated those ten use cases and ranked them using Slido to determine which ones were most important for the OBIS community. Of the ten use cases that were selected by the project team, eDNA was identified as most important (Figure 1).



Which use cases are most important for the OBIS community to tackle?

1. eDNA
2. Environmental and community measurements
3. Specimens with media
4. Ecological survey data
5. Taxonomic checklist
6. Camera trap
7. Tissue samples
8. iNaturalist
9. Taxonomic treatments in literature
10. Automatic moth traps

Figure 1. Screenshot of Slido poll used to determine which use cases that were proposed by GBIF (at the time of the poll) were the ones the OBIS community would consider most important.

Many potential datasets were identified to work on for this project team (see Appendix I), but not all were practical, and in the end only a few received significant attention and are therefore addressed in this report. Early on we identified the ARMS / eDNA dataset managed by EurOBIS as being a particularly complex use case that would serve as the guiding one for the project team, but there were difficulties in sharing the data.

Absence Use Case

Accurate documentation of where species occur and how their distributions change over time is crucial for understanding and predicting population trends and movements. This information is the raison d'etre of biodiversity data aggregators such as GBIF and OBIS and forms the foundation for ecological and conservation studies, allowing researchers to analyze how species are responding to environmental changes.

To model population trends accurately, the information about when a species is present is as important as the information about when a species is NOT present, traditionally known as "absent". The Darwin Core standard provides a space to document both presences and absences using the occurrenceStatus term. However, providing compelling evidence for a taxon being absent is not straightforward since there is not yet consensus about the exact definition of what an absence is, resulting in ambiguous documentation for absence data in Darwin Core and a heterogeneous usage of the standard.

Acknowledging the relevance of documenting absence data in a standardized manner, the OBIS GUMPT got in touch with the GBIF Discourse community, which led to the creation of a subteam composed of both OBIS and non-OBIS members (The Absence Use Case team). This subteam met once a month from December 2022 to July 2023 to identify, name and define all the different types of absences used by the biodiversity scientific community. Eventually providing the following standardized list of terms for absence data:

- True Absence
- Non-detection
- Background points
- Reporting completeness

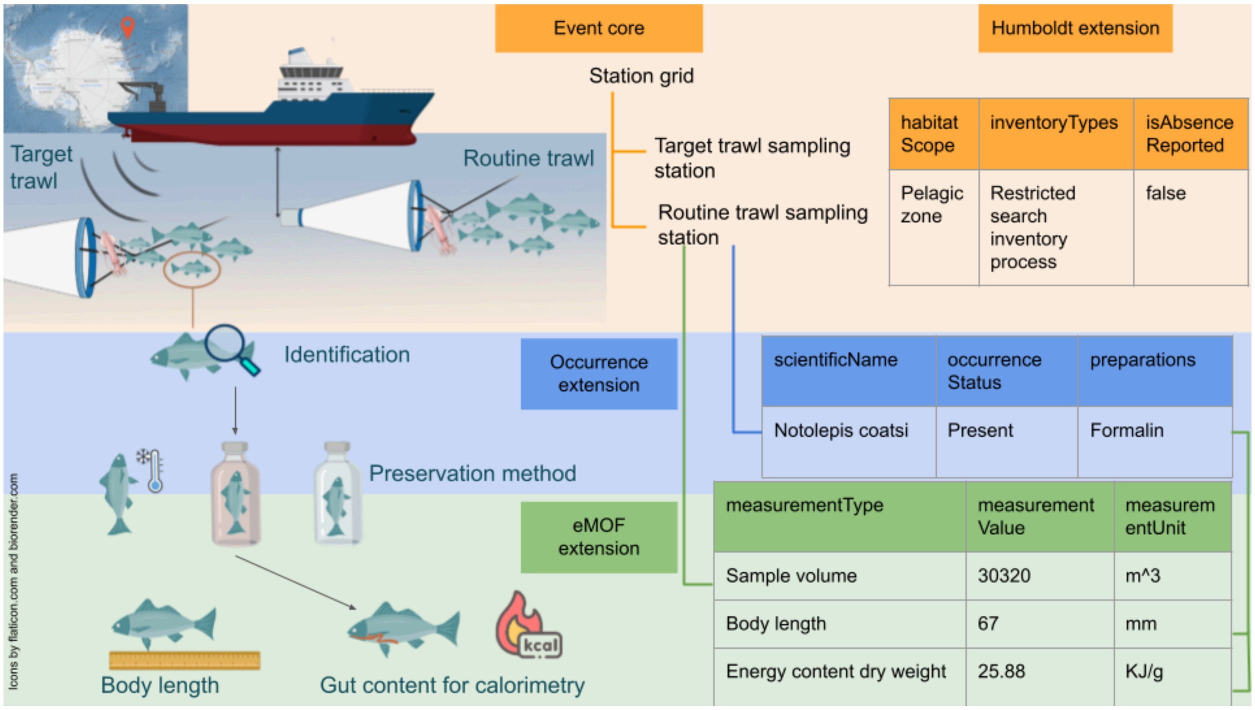
The work of this subteam continues as they further refine the definitions with the ultimate goal of publishing a paper for these definitions. A few of the team members in this group were also part of the Humboldt Extension Task Group and actively connected the work of these two groups.

Collaboration with the Humboldt Extension Task Group

The lack of reporting standards for species inventories hinders their utility for biodiversity assessment. While the Darwin Core standard covers some inventory information, the Humboldt Extension Task Group has developed a Humboldt Extension (a Darwin Core extension for the Event Core) to address these limitations, to provide more comprehensive data sharing and integration.

The main objective of the Humboldt Extension is to allow data users to be able to make inferences of absences from an ecological species inventory dataset. The types of absence (true absence or non-detection) referred herein can be inferred based on the completeness of the sampling and reporting. Humboldt Extension terms that describe the scope of an inventory such as targetTaxonomicScope and completeness of reporting such as isTaxonomicScopeFullyReported are helpful to infer non-detection of target taxa. Indication of true absence of target taxa can be evaluated using the scope terms and reporting completeness terms, as well as taxonCompletenessReported and taxonCompletenessProtocols.

Yi-Ming Gan has attempted to map an AntOBIS dataset (referred as BROKE-West dataset herein) with the Humboldt Extension in a Darwin Core Archive: https://ipt.gbif.org/resource?r=brokewest-fish



habitatScope	inventoryTypes	isAbsenceReported
Pelagic zone	Restricted search inventory process	false

scientificName	occurrenceStatus	preparations
Notolepis coatsi	Present	Formalin

measurementType	measurementValue	measurementUnit
Sample volume	30320	m^3
Body length	67	mm
Energy content dry weight	25.88	KJ/g

Figure 2. Diagram illustrating how various data components of the BROKE-West dataset are mapped to the Event Core, Humboldt Extension, Occurrence Extension and extended Measurement or Fact (eMoF) Extension.

Yi-Ming also attempted to infer non-detection of the target taxa of a presence-only dataset using the Humboldt Extension in the GitHub repository:
https://github.com/biodiversity-aq/humboldt-for-eco-survey-data
The purpose of doing this mapping was to document the limitations imposed by Darwin Core and provide that as feedback to refine the Humboldt Extension and the development of the Unified Model. The key findings relevant for GUMPT are described below.

Scopes are difficult to be expressed due to the limitation of the Darwin Core star schema

The scopes related terms such as taxonomic scope and organismal scope are difficult to express when flattened into one single table in a Humboldt Extension extending the Event Core. The scopes information has to be duplicated for all Events for which they are in effect. This could be more efficient if it could be a foreign key to a target table. However, the limitation of a star schema prevented this (because relationships between an extension and the core can only go one level deep). This has been provided as feedback and hopefully can be resolved with the new conceptual model.

Current Humboldt Extension lacks identifiers field

Even though identifiers for each Humboldt term can be used with the dwciri: namespace the schemas used to map data to Darwin Core and Extension terms do not include the terms from the dwciri: (nor ecoiri: namespace). The Integrated Publishing Toolkit (IPT), which is the most popular tool to publish biodiversity data as Darwin Core Archives, relies on these schemas to allow datasets to be mapped. The lack of identifiers for the terms can have some drawbacks. For instance, targetTaxonomicScope is currently only accepting scientific names instead of identifiers such as life science identifiers (LSID). This can lead to ambiguity if a name is a homonym, and work is underway to address this in GBIF (https://github.com/gbif/pipelines/issues/217). The identifiers common model of the new conceptual model may additionally help to mitigate this issue.

The objective of exploring how absence of detection can be represented in the new conceptual model has not been achieved as the publishing model of the Humboldt use case is not yet available and will be developed once the Humboldt Extension terms are ratified.

Baited Remote Underwater Vehicle to CamtrapDP

A case study to understand mapping Baited Remote Underwater Video Stations (BRUVS) data to Camera Trap Data Package (Camtrap DP) was attempted by Sachit Rajbhandari. This work was done with the knowledge of ongoing activities (Webinar on Exploring camera-trap data) and aligns to a new technical guide (Best Practices for Managing and Publishing Camera Trap Data) released by GBIF for community peer review.

Camtrap DP is a community-developed data exchange format using Frictionless Data Package. The Camtrap DP aims to make sharing camera trap data more accessible and standardized. The BRUVS observation data from 5 surveys at the Ningaloo Reef, Western Australia in 2019 were used for this case study. It is a part of the Marine Biodiversity Hub D3 project and downloaded from GlobalArchive. The video and images from the BRUV surveys can be downloaded from the CSIRO Data Access Portal.

Along with the data structure, metadata is also included, which holds essential information about the entire dataset. The Camtrap DP dataset consists of the following files (datapackage.json, deployment.csv, media.csv and observations.csv), which were mapped to BRUVS data. A Python script was written to perform the mapping between the source dataset files and Camtrap DP files.

The datapackage.json file is created manually for this case study as the dataset is not loaded into a data management platform/tools such as Agouti or TRAPPER or CamtrapR, which feature exporting data into Camtrap DP. Later, after a recommendation from Peter Desmet, the datapackage.json file was created using a new release candidate of its Integrated Publishing Toolkit (IPT) for testing and feedback.

The observations.csv file serves to store essential information such as the taxonID and scientificName. However, it's important to note that the detailed taxonomic information is not included at the observation level within this CSV file. Instead, the taxonomic details are captured and stored in the datapackage.json file.

In this dataset, the EventMeasure (Stereo) tool has been used to annotate, record events and report abundance. Stereo measurement allows measurement of 3D position relative to the camera system along with 3D length measurement, including the range and pose of the measured object. The observations.csv covers taxon, count, life stage, sex, behaviour and/or individual, but is unable to store other measurements such length or weight.

In Camtrap DP 0.6, observations were split into media-observations and event-observations. When media files are initially grouped into events or sequences, and then observations are generated based on these events, the "event-observations" approach is used. This method allows for a higher-level perspective of data organization, often suitable when events carry more significance than individual media files. Conversely, if each media file has been individually assessed without being part of larger events, the "media-observation" approach is employed. This method offers a granular level of detail and is particularly valuable for tasks such as training image recognition models, where the focus is on individual media files. With Camtrap DP 1.0 RC, observationLevel is introduced, which can be media or event.

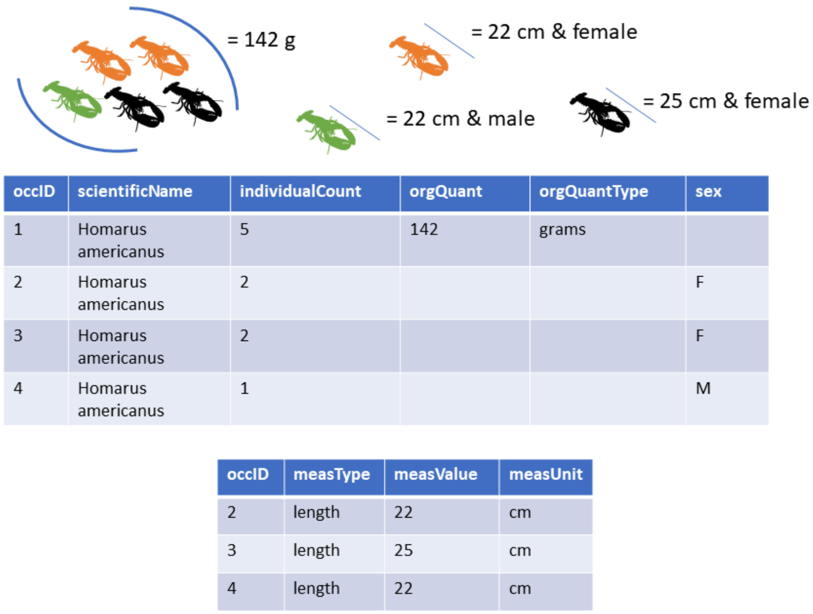
Camtrap DP data mapping to the new conceptual model

A Camtrap DP is a simpler publishing data model whereas a conceptual data model is a high-level representation of camera trap data covering entities and relationships that represent concepts such as **Organism** and **Occurrence**. Sachit Rajbhandari looked into mapping the Camtrap DP dataset to the GBIF conceptual model following the recommended guidelines. The examples from the GBIF model-tests GitHub repository were helpful during the mapping process. For mapping, a Postgresql database with the new conceptual model structure was created using the script schema.sql and created tables were populated from the source data using a Python script.

A camera deployment is the primary **Event** followed by other *event_type* such as Image capture and observation. An **Event** occurred at a particular *Location* with special assertions **Georeference** providing lat/lon and other geo-information of the events. The captured video file is represented as **DigitalEntity** and the organism observed is represented by **MaterialEntity** *concept*. An Entity record is created for both **DigitalEntity** and **MaterialEntity** with *entityType* specified. The observed **Organisms** have *Assertions* about the count of individuals, lifeStage, sex, behaviour, etc. defined as *assertion_type*. Each **Organism** is given **Taxon Identifications** by experts with the ScientificName assigned. To keep the mapping process simpler, ***Agents, Assertions, Citations***, and **Identifiers** for **Agents**, **References** and **Protocols** were skipped.

Maine Inshore Trawl Survey to Conceptual Model

One issue that is not currently handled by Darwin Core or the EMoF Extension is when a group of organisms has measurements but individuals from the group also have measurements. One dataset that this occurs for is the Maine Inshore Trawl Survey, which is currently shared with OBIS and GBIF. This is a trawl survey that occurs twice per year and uses a stratified random sampling scheme for selecting sites. The Maine Inshore Trawl Survey provides a weight for the entire catch of a species, but also length measurements for individuals in that catch. Currently you would need one row in the occurrence file for the entire catch and then multiple additional rows for each group of fish of the same length. This leads to the potential for double counting of the individualCount for a species (Figure 3).



occID	scientificName	individualCount	orgQuant	orgQuantType	sex
1	Homarus americanus	5	142	grams	
2	Homarus americanus	2			F
3	Homarus americanus	2			F
4	Homarus americanus	1			M

occID	measType	measValue	measUnit
2	length	22	cm
3	length	25	cm
4	length	22	cm

Figure 3. Diagram showing an example catch for the Maine Inshore Trawl Survey where the potential for double counting individuals is possible given the current capabilities of Darwin Core.

As part of the work of this project team, Abby Benson attempted to align the raw Maine Inshore Trawl Survey data to the conceptual model using instructions provided to the collections conceptual model pilot project. While she was able to make significant progress the data were never finalized. Even though the conceptual model is best for theoretically modeling data, a publishing model would have been easier to work with. One issue with the conceptual model is that it's composed of many different tables and it was sometimes difficult to figure out which ones needed to be created before others and all required unique identifiers to be created and then used in related tables. Since we never reached a completed state on this dataset, it wasn't possible to do a full assessment of how well the new conceptual model would have fit this dataset.

Future Directions

While this project team addressed most of the tasks set out at the beginning, they were not always fully completed by the end of the project. The new conceptual model is still in development and continuing to undergo changes. The conceptual model can be difficult to map data to but most of the publishing models are still in development. A future team could work with those publishing models once they are completed and make a fuller assessment. A deeper look at using Frictionless Data Packages would also be beneficial.

Appendix I

List of datasets originally identified for the project team to potentially work on. The ones highlighted in yellow were identified to be worked on by the project team.

1. eDNA barcoding
    a. eDNA dataset with abiotic measurements, EurOBIS, Ruben
    b. Metabarcoding Lab, OBIS Colombia, Martha
    c. UNESCO eDNA expeditions, Ward
    d. PacMAN, OBIS SWP, Kevin M. and Ward
    e. Marine microbes eDNA datasets from national reference stations, OBIS-AU, Sachit
    f. MBON eDNA dataset, OBIS-USA, Abby
    g. eDNA Antarctic Lakes dataset, OBIS Antarctica, Ming and Andre Heughebaert
    h. Rockefeller University dataset - Abby
2. Camera trap
    a. Video and still camera transects, OBIS SWP, Kevin M.
    b. Video transects along new deep sea MPA areas, OBIS Colombia, Martha
    c. DFO Video transects, OBIS Canada, Maria
    d. BRUV dataset, OBIS-AU, Sachit
    e. MBARI VARS, OBIS-USA, Abby
    f. IFCB, OBIS-USA, Abby
3. Tissue samples
    a. “Preserving and sharing the marine genetic diversity of Colombia through the tissue collection of the Marine Natural History Museum of Colombia - Makuriwa of INVEMAR”, OBIS Colombia, Martha
4. Automatic moth traps
    a. ARMS, EurOBIS/Secretariat (in OBIS use case)
5. Global malaise programme
6. iNaturalist
    a. Wembury Bioblitz, OBIS UK, Kevin P.
    b. eOceans, Elizabeth
    c. Sea lion observation reports, OBIS Colombia, Martha
    d. NatureWatchNZ, OBIS SWP, Kevin M.
7. Specimens with media
    a. CSIRO Marine Invertebrates Image Collection (MIIC), OBIS-AU, Sachit
8. Environmental and community measurements
    a. Nansen Legacy data, Antarctic OBIS, Ming
    b. Typical historical dataset, OBIS Historical, John
    c. Maine DMR Trawl, OBIS-USA, Abby
    d. NZ research trawl survey, OBIS SWP, Kevin M.
    e. Trawl (and other gear) fishery research survey datasets, OBIS Canada, Maria
9. Taxonomic treatments in literature
    a. WoRMS?
10. Malaise trapping for reference barcode collection
11. Taxonomic checklist

    a. Offshore and coastal hydrocarbon sensitivity maps, OBIS Colombia, Martha

12. Ecological survey data exchange specification
    a. LTER dataset, EcoComDP model, OBIS-USA, Abby
13. Biotic interactions
14. Missing Use Cases
    a. SCAR Southern Ocean Diet and Energetics Database, Antarctic OBIS, Ming
    b. Primary literature sources for historical data (unless taxonomic treatments in literature is most appropriate?), OBIS Historical, John
    c. Habitat data like seagrass, Ward I think Ward means adding habitat type as measurementValue and measurementValueID in eMOF using eg. https://eunis.eea.europa.eu/habitats-names.jsp - see seagrass data schema https://docs.google.com/spreadsheets/d/12LRma67Nwl54eT41HpDAYfOCqXvGKKInPPCSBonTQIg/edit#gid=975591799