Biological Data Work Flows

O&A Information and Data Centre

Exported on 04/23/2019

Table of Contents

1 Author:

Dave Watts, CSIRO Oceans and Atmosphere Data Centre and OBIS Australia Node manager (http://www.obis.org.au)

Katherine Tattersall, CSIRO Oceans and Atmosphere Data Centre

2 Purpose of this document:

There are four primary biological data aggregators or portals that acquire and publish marine biological data in Australia.

They are:

OBIS - Ocean Biogeographic Information System http://www.iobis.org[1] (New site in development https://portal.obis.org)

ALA - Atlas of Living Australia http://www.ala.org.au

GBIF - Global Biodiversity Information Facility http://www.gbif.org

AODN - Australian Ocean Data Network http://www.aodn.org.au

This document describes best practice guidelines for loading data into the most relevant portal for different types of biological data, with the capacity to move that data into any of the other portals in a seamless and robust manner. As transformation of data does not necessarily capture all the nuances of the original data, it is important that all portals reference the prime or point of truth data*set.*

1 http://www.iobis.org

3 Characteristics of each portal:

3.1 OBIS - Ocean Biogeographic Information System (International)

The OBIS aim is acquire species occurrences records from a well-defined network of OBIS-endorsed data providers. It is the role of the endorsed data providers to ensure the data being published is of good quality and is marked up with locations and dates. It is expected that all taxa are matched to the World Register of Marine Species (WoRMS).

The current harvesting implementation uses the Integrated Publishing Toolkit (IPT) to link data structures to the Darwin Core standards: occurrences; extendedMeasurementorFact; media; events ... etc. The last three schemas allow richer data types and environmental data to be published and OBIS is working to expose these data in their new portal. As an example of publishing non-ascii data, Images from the CSIRO Marine Invertebrates Image Catalogue are now presented in WoRMS from records using the media schema.

3.1.1 Submitting data to OBIS-AU

3.1.1.1 OBIS-AU: The Australian OBIS node

The OBIS Australia Node webpage (http://www.obis.org.au) gives more information about how to publish data. The agreed pathway for Australian marine data to publish to OBIS is via the IPT, hosted and managed at CSIRO by the OBIS-AU. It is important to note that there is no alternative pathway to deliver data to OBIS other than hosting it on an OBIS endorsed IPT.

Once data is made public on the IPT it is immediately harvested by OBIS. If a dataset is replaced or updated on the IPT then OBIS will completely replace a dataset. This means that there is no issue of stale or inconsistent data at OBIS versus the published data at the IPT.

3.1.1.2 Australian Antarctic Node

A second IPT node in Australia is the recently established Australian Antarctic Data Centre (AADC) IPT (http://data.aad.gov.au/ipt). This node will be directly harvested by OBIS and potentially direct to ALA. Once the data is in ALA, the data will be visible at AODN. The AADC will in the first instance handle all Australian Antarctic Program data. If there is any contention about the jurisdiction of the data, OBISAU and AADC will resolve it as it is important that data is not needlessly duplicated within any network or portal.

3.1.2 Identifying marine data in OBIS

Any non-marine taxa records are ignored by OBIS. Unaccepted marine taxa names (synonyms, non-standard names etc.) can be used and OBIS will use WoRMS to report on data for both the accepted name and any unaccepted names. OBIS reports taxa names that require correction to the IPT data provider.

3.1.3 Metadata and data format standards

OBIS uses the EML and DwC standards (see Metadata, Tools and Data *Standards* section below).

3.2 GBIF - Global Biodiversity Information Facility (International)

GBIF is the largest global biological data aggregator and takes taxa records from any domain, including marine but also terrestrial and freshwater aquatic species.

3.2.1 Submitting data to GBIF

GBIF, like OBIS, harvests records from IPT instances. Some of the IPT instances are the same as those that are harvested by OBIS, but not all. To be harvested by GBIF, each dataset within the IPT must be registered with GBIF using the 'registyer' btutton in the IPT. Unlike OBIS there is little if any feedback on issues of taxa, date or spatial errors to IPT managers.

3.2.2 Identifying marine data in GBIF

There is no obvious filter for marine taxa records at GBIF.

3.2.3 Metadata and data format standards

GBIF uses the EML and DwC standards (see Metadata, Tools and Data Standards section below).

3.3 ALA - Atlas of Living Australia (National)

ALA is a collaborative, national project that aggregates Australian biodiversity data from multiple sources and makes it freely available and usable online. ALA is the Australian node of the Global Biodiversity Facility GBIF[2]. ALA takes taxa records from any domain, including marine but also terrestrial and freshwater aquatic species. ALA also hosts a separate animal tracking data portal, ZoaTrack[3]. Data from both portals can be downloaded as plain csv or in a format consistent with Darwin Core[4].

3.3.1 Submitting data to ALA

ALA has standard spreadsheets (available here[5]).

3.3.2 Identifying marine data in ALA

ALA tags taxa with a marine attribute and there are web service calls to retrieve datasets and records containing marine taxa.

3.3.3 Metadata and data format standards

ALA publishes data using the DwC standards (see Metadata, Tools and Data Standards section below).

2 http://confluence.csiro.au/www.gbif.org
3 https://zoatrack.org/
4 http://rs.tdwg.org/dwc/
5 https://www.ala.org.au/submit-dataset-to-ala/

3.4 AODN - Australian Ocean Data Network (National)

The AODN Portal provides access to Australian marine and climate science data, including biological data but also chemical and physical properties of the ocean and atmosphere. AODN uses a different approach for publishing data. For biological data, it typically uses data delivery services from GeoServer instances for end users to filter and download (e.g. as a csv file), either from a GeoServer managed by AODN or from other nodes within the AODN network. THREDDS servers are used to deliver data for primarily physical datasets.

3.4.1 Submitting data to AODN

IMOS data is managed and loaded into AODN. For other sources of data, please contact AODN to facilitate submission on a case-by-case basis

3.4.2 Metadata and data format standards

AODN uses the WFS OGC standard to publish data from a geospatial data server. The only forced conformance requirement is the geometry object (point, line or polygon) with all other data elements *not forced to be conform to any particular controlled vocabularies. There are AODN guidelines for setting up data sources.*

The format of the WFS sourced csv downloads are not controlled in the same manner as an IPT dataset or as a download from OBIS, GBIF or ALA. In many instances, the naming convention of the elements within the csv data is adhoc making integration with similar data types an issue.

In the AODN there is no indication of how much data is available through a Geoserver data service or if there have been any changes, additions or deletions from prior downloads. It is not certain if the FID field within a WFS dataset is stable and best practice would be to include an immutable primary key identifier that can be used not matter what type of publication mechanism is utilised

.

4 Metadata, Tools and Data Standards:

4.1 EML - Ecology Markup Language

EML is a metadata standard for describing biological datasets. There is no simple process to programatically convert ISO metadata to EML. As EML is relatively simple and the IPT editor is mature, it is easy to manually transcribe content from an ISO metadata record.

4.2 IPT - Integrated Publishing Toolkit

Built by GBIF as a means to handle very large datasets over limited bandwidth connections. The manual is available from https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki

IPT is a web-based application that packages biological data into a self-contained zipped file for download. The IPT has a metadata authoring tool to create dataset metadata records in Ecological Markup Language (EML). This file contains the EML metadata along with well-marked CSV data conforming to the DwC vocabularies. The IPT forces dataset authors to match internal data identifiers/database columns to the DwC terms. There is no scope for a user to define their own column names.

The IPT builds a snapshot of the dataset (as a zipped file) and marks it with an incrementing version number. If any of the underlying data is changed, then a new snapshot needs to be published before data aggregators can see any changes.

4.3 WoRMS - World Register of Marine Species

WoRMS (http://www.marinespecies.org[6]) is an authoritative and comprehensive list of names of marine organisms, including information on synonyms. While the highest priority by the taxonomic editors goes to valid names, other names in use are included *(synonyms, deprecated names, vernacular names)* so that this register can serve as a guide to interpret taxonomic literature. There are webservices to query WoRMS[7] and it uses near fuzzy matching algorithm to match names which is particularly useful for misspelled scientific names.

4.4 DwC - Darwin Core

DwC (https://dwc.tdwg.org/) is a set of terms used by the biological community (and built into the IPT) to publish data in a well-defined standard format (Wieczorek et al. 2012)

The following DarwinCore fields are the miniumum expected to publish data to OBIS:

- occurrenceId - this is the primary key or unique identifier for a row of occurrence data. It is best constructed at source and should stay immutable for the lifetime of the data.
- occurrenceStatus - values of 'present' or 'absent'
- BasisofRecord - values are HumanObservation, MachineObservation, Sample, PreservedSpecimen, FossilSpecimen and LivingSpecimen. The IPT enforces this strict set of values for data to be published.
- decimalLatitude - typically using and assumed to be WGS84 (EPSG:4326)

6 http://www.marinespecies.org/
7 http://www.marinespecies.org/aphia.php?p=webservice

- decimalLongitude -
- eventdate - date of occurrence in ISO data format (e.g. 1977-01-19 23:45). It is preferred that this date is in UTC.
- ScientificName - the scientific name of the taxa
- ScientificNameId - OBIS requires the URN of that taxa from the World Register of Marine Species: WoRMS[8] to be included where possible. For example, the Wandering Albatross (Diomedia exulans) has an id of 212583 (see http://www.marinespecies.org/aphia.php?p=taxdetails&id=212583). The URN for this species is urn:lsid:marinespecies.org[9]:taxname:212583. If the name is not in WoRMS, then identifiers from the Australian Faunal Directory AFD is suitable.

There are also extensions available to use within DwC and the most useful is EMoF - Extended Measurements or Facts - for adding unlimited extra details about biological data which can be related to the occurrence or an event. De Pooter et al (2017) describes EMoF and its implementation for OBIS data.

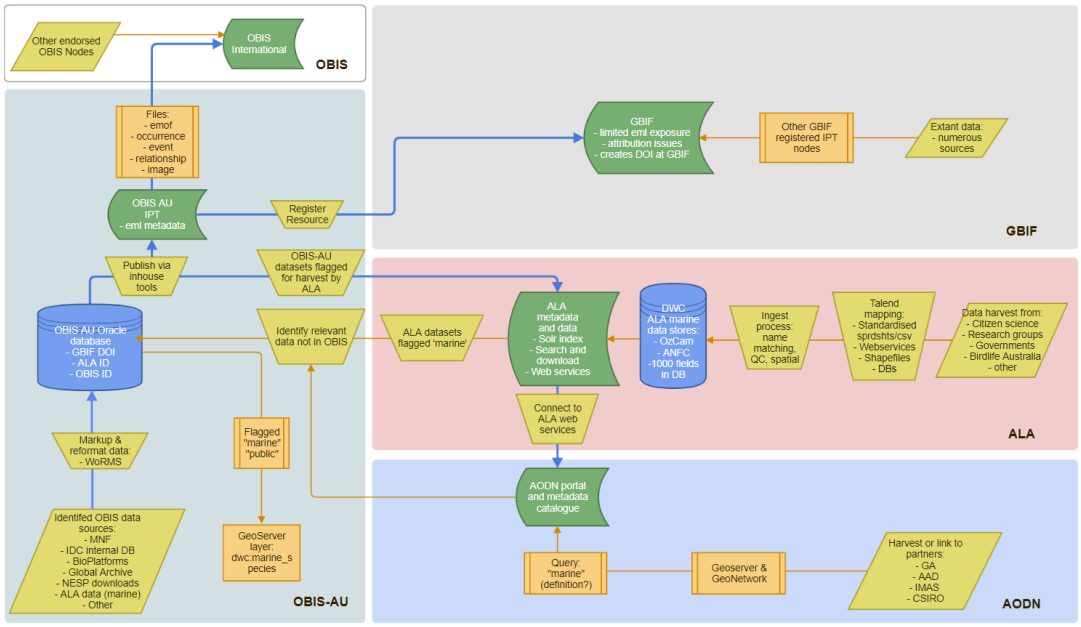
8 http://www.marinespecies.org/
9 http://marinespecies.org

5 Publishing Australian Marine Biological Data:

5.1 Existing data flows:

The following diagram shows the potential data flows within the ALA, GBIF, AODN and OBISAU network. Data can be lodged at any portal but there are different types of work required to migrate the data to any of the other portals. The preferred data flow is in blue.

Not shown are web service linkages for discovering new datasets between OBISAU and ALA and the AODN using ALA web services to show species occurrence data within the AODN portal.



NOTE - If your data is derived from an Australian Antarctic Program project, please contact the AADC. If there is any confusion about who should manage the data given the overlapping jurisdiction of the Southern Ocean, OBISAU will liaise with the AADC.

5.2 Submitting data:

The following are descriptions of how the data can be lodged with a portal for publication. There are various actors that may be involved.

User - the person providing new or revised data to a portal.

OBISAU - the team managing the OBISAU IPT and underlying data validation and publishing mechanisms.

ALA - staff at ALA managing the ingesting, publication and validation of ALA data.

A) If data is to be lodged with OBIS Australia:

User: Inform OBISAU of a new dataset using the email address OBISAU@csiro.au.[10]

OBISAU - Will review the data and if required suggest improvements before loading the data and metadata into the OBISAU IPT. If the dataset is suitable for but not at ALA, OBISAU will markup the dataset metadata so that ALA can be informed of a new dataset via an OBISAU webservice. Once published at ALA, it will appear in the AODN Portal. If the data is not at GBIF, then the dataset is 'registered' with GBIF in the OBIS IPT and is then immediately harvested by GBIF.

B) If data is lodged with ALA:

User: Typically users will uploaded data to ALA using standard spreadsheets (available here[11]).

ALA: The data will be loaded and a unique dataset or collection identifier created (UID e.g. dr2345 or co961). Once validated and the taxa have been tested as 'marine', it is visible at AODN within the 'Data collection' WHAT NAME?? using the ALA web services.

OBISAU: ALA has many web services and it is possible to identify datasets containing one or more records of marine taxa (see Appendix for details). OBISAU has vetted 400+ marine datasets from ALA and uses this web service to identify potential new datasets. Some of the datasets do not have locations or dates and are excluded from harvesting for OBIS.

The occurrence records from a dataset can be retrieved (see Appendix for details) as JSON. The JSON response is parsed and update a database table for new or changed records. Each dataset requires an IPT resource name and OBISAU uses the convention of appending the ALA UID to the prefix 'ala_' (example is 'ala_dr2345'. Taxa are matched to WoRMS using its web services and an EML metadata record is authored using the details in the home page of the dataset at ALA.

C) If data is lodged with GBIF:

User: Typically users will upload data to GBIF using an IPT. Please inform OBISAU of a potential dataset using the email address OBISAU@csiro.au.[12]

OBISAU: Harvest the GBIF data, check if suitable and add to the OBISAU IPT. If the dataset is suitable for but not at ALA, OBISAU will markup the dataset metadata so that ALA can be informed of a new dataset via an OBISAU webservice. Once published at ALA, it will appear in the AODN Portal.

D) If data is lodged with AODN:

User: If your data has been published via an OGC WFS server, please inform OBISAU of a potential dataset mentioning in particular the metadata URL.

OBISAU: If suitable, the WFS data will be downloaded, any taxa will be matched to WoRMS and every field will be mapped or combined to a relevant DwC term and loaded into the local OBISAU database. An EML metadata record will be created using the AODN metadata record as a basis. Links will be provided to the original data source. If the dataset is suitable for but not at ALA, OBISAU will markup the dataset metadata so that ALA can be informed of a new

10 mailto:OBISAU@csiro.au.
11 https://www.ala.org.au/submit-dataset-to-ala/
12 mailto:OBISAU@csiro.au.

dataset. If the data is not at GBIF, then the dataset is 'registered' with GBIF in the OBIS IPT and is then immediately harvested by GBIF.

Finally, the best practice is to use Option A as data can be moved easily to ALA, AODN and/or GBIF. It allows OBISAU to mark up the data to the OBIS requirements. The preferred pathway for moving the data is shown as a thick blue arrowed line.

5.3 References

De Pooter D, Appeltans W, Bailly N, Bristol S, Deneudt K, Eliezer M, Fujioka E, Giorgetti A, Goldstein P, Lewis M, Lipizer M, Mackay K, Marin M, Moncoiffé G, Nikolopoulou S, Provoost P, Rauch S, Roubicek A, Torres C, van de Putte A, Vandepitte L, Vanhoorne B, Vinci M, Wambiji N, Watts D, Klein Salas E, Hernandez F (2017) Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. https://doi.org/10.3897/BDJ.5.e10989

Wieczorek, John; D. Bloom; R. Guralnick; S. Blum; M. Döring; R. De Giovanni; T. Robertson; D. Vieglais (2012). Darwin Core: An Evolving Community-developed Biodiversity Data Standard.. PLoS ONE. 7 (1). PMC 3253084? . PMID 22238640. doi:10.1371/journal.pone.0029715.

5.4 Appendix

ALA Web services

Purpose	Details
Identification of marine datasets at ALA	https://biocache.ala.org.au/ws/occurrences/search?q=species_habitats:Marine&facets=facet&pageSize=0&flimit=1000 (where [facet has values of either collection_uid or data_resource_uid) will retrieve marine datasets (defined as having one or more records of marine taxa). flimit limits the number of faceted records returned startIndex is the starting number of records offset from 1 The returned JSON is an array containing dataset name, dataset identifier, number of marine records

Purpose	Details
Retrieving records from a dataset	https://biocache.ala.org.au/ws/occurrences/search?q=species_habitats:Marine&fq=data_resource_uid:uid where uid is the ALA dataset identifier (e.g. 'dr2345' or 'co88') The returned JSON is an array with mostly DarwinCore terms. You can set a download limit or starting point within the recordset using pageSize and startIndex parameters. pageSize is the maximum number of records retrieved per request Example is https://biocache.ala.org.au/ws/occurrences/search?q=species_habitats:Marine&facet=no&fq=data_resource_uid:dr734&pageSize=500&startIndex=501[13]

13 https://biocache.ala.org.au/ws/occurrences/search?q=species_habitats:Marine&facet=no&fq=data_resource_uid:dr734&pageSize=500000&startIndex=1