Marine Data Hub Data Management and Publishing Guidelines

INSTITUTE OF MARINE AFFAIRSS

Marine Data Hub Data Management and Data Publishing Guidelines

Hilltop Lane, Chaguaramas
Trinidad & Tobago

Revised on 22 June 2023

Contents

The Global Biodiversity Information Facility

Chapter

1

What is the Global Biodiversity Information Facility?

"The Global Biodiversity Information Facility (GBIF) is an international network of country and organizational participants that exists to enable free and open access to biodiversity data from all sources and to support biodiversity science, environmental research, and evidence-based decision-making. GBIF operates as a federated system of distributed data publishing efforts, coordinated through a global informatics infrastructure and collaborative network." -GBIF Secretariat (2021)

- An introduction to GBIF

What is Darwin Core?

Darwin Core is a standard maintained by the Darwin Core maintenance group within GBIF. It includes a glossary of terms intended to facilitate the cohesive sharing of information regarding biodiversity data by providing standardized identifiers, labels, and definitions under which the data is shared. Darwin Core (DwC) is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and other related biodiversity information.

The Darwin Core Standard is made up of a set of standardized data entry headers- known as cores or DwC's- with protocols guiding the format in which the data under these headers are maintained. The Maintenance Group, which oversees the modification and enhancements of these standards, have created baseline cores and subsequent extension cores that can be used in data entry. The cores used are dependent on the data that is being captured and can be chosen by the organization as needed.

The Marine Reference Collection

Chapter

2

The Marine Reference Collection (MRC) at the Institute of Marine Affairs (IMA) houses specimens from a wide array of scientific projects and collection exercises. As such, the MRC is a prime example of biodiversity data that can be digitized, standardized, and uploaded for public access.

When determining the standardization protocols to be used, the first steps include finding the GBIF data entry template that best fits the information available. The three most common templates used include:

- ☑ The event_ipt_Template- generally used for recording sampling events.
- ☑ The occurrence_ipt_Template- generally used for recording species data based on locality and often includes coordinate data for mapping purposes.
- ☑ The taxon_core_template- generally used for classification of specimens based on taxa.

The above templates are simply guides and can be edited to include more information as is pertinent to the dataset. In the case of the MRC, the dataset required further DwC terms to reflect its spatial information. As such, the standardized data entry sheet for the MRC mainly utilized the occurrence template with the addition of a few spatial cores from the Darwin Core Maintenance Group (2021).

DARWIN CORE (DWC) TERMS

The Darwin Core standard and is maintained by the Darwin Core Maintenance Group. The standard comprises a glossary of terms, the intention to facilitate the sharing of information about biological diversity through the use of definitions, identifiers and labels.

A list of the core terminology is available at List of Darwin Core terms - Darwin Core (tdwg.org). The following are so of the popular terms used within the MRC.

- occurrenceID
- basisofRecord
- otherCatalogNumber
- individualCount
- year
- month
- day
- eventDate
- eventRemarks
- verbatimDepth
- minimumDepthinMeters
- maximumDepthinMeters
- habitat
- samplingProtocol
- preparations
- verbatimIdentification
- scientificName
- Taxon (the terms listed below are accepted)
    - kingdom
    - phylum
    - class
    - order
    - family
    - genus

- recordedBy
- verbatimLocality
- locality
- country
- countryCode

The specimens in the MRC date as far back as the 1970's. As such, data quality checks were required to be performed on the specimen information cards attached to ensure that it reflected the most up-to-date taxonomic information. As part of this process, the species names and family information of each species was double checked using the World Register of Marine Species (WoRMS) website.

The following links provide detail guidance on data preparation and data quality which is a critical step in preparing data for publication. The information is detailed in the Integrated Publishing Tool (IPT) User Manual. Please consult the links to obtain a full understanding of the related aspects of data quality management.

- Darwin Core Archives – How-to Guide
- Data Preparation :: GBIF IPT User Manual
- Data Quality Checklist :: GBIF IPT User Manual
- GBIF Metadata Profile – How-to Guide :: GBIF IPT User Manual

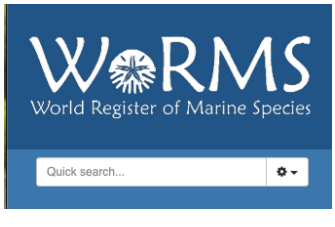
Taxonomic Validation

Taxonomic Validation may be achieved via GBIF or WoRMS Application Programming Interface (API) and Species Matching Routine and using the Species Matching tool after validating the original data, and loaded into OpenRefine. The links to the different approaches are listed below.

- Taxonomic validation with the GBIF API
- Taxonomic validation with the WoRMS API
- Taxonomic validation with GBIF Species Matching
- Taxonomic validation with manual search

The manual search approach is listed below.

Step 1: A quick search of each of the species' names listed on the identification cards were performed on the website.



Step 2: Once the exact match was found, ensure that the species name has a status of "Accepted" as seen below

Step 3: Once the status was accepted, the species' information is then used to construct the occurrenceID (a DwC term that represents a unique identifier for the specimen).

The occurrenceID followed the following naming convention format:

"IMA-MRC-ACT-"*ORDER*"-"*FAMILY*"-*"1, 2, 3…"*

where

"Order" represents the taxonomic order listed on the WoRMS website,

"Family" represents the taxonomic family listed on the WoRMS website and

the numbers (1,2,3,etc) represented the jar number in the collection. For example, if there are five jars in the Family Lutjanidae, the occurrenceID's would read "…. – Lutjanidae-1" , "…. – Lutjanidae-2", "…. – Lutjanidae-3", etc

ACCESSING THE MRC & ADDITIONAL INFORMATION

To access the MRC:

1) The IMA MRC file has been standardized to Darwin Core Standards, listed

IMA MRC.csv

here.

Formatting of the file:

1) The file must remain as a .csv with UTF coding. This is to ensure readability on GBIF''s IPT.

MRC SPECIMEN DIGITIZATION

1. Print and cut ID cards on water proof paper
2. Dry specimen in trays and using paper towels
3. Prepare the work area to pin specimens down
4. Place rulers both vertically and horizontally
5. Pin specimen down in the desired orientation
6. Count/record the specimens present in each jar
7. Photograph whole specimen with ID
8. Photograph whole specimen with ID and description card
9. Photograph close up image for smaller specimens/distinguishable features/unique features etc
10. Place specimen back with ID and description card

*Use digital editing software such as Adobe Light Room, Sketchbook or equivalent to edit/enhance/clean images

11. Naming convention must follow:

"IMA-MRC-ACT-"*ORDER*"-"*FAMILY*"-*"1, 2, 3…"*

12. Ensure that lids are seals properly and not leaking. Use cling wrap to secure/stop leaks.

Chapter 3

Historical Data Sets

The Institute of Marine Affairs has been operational since 1974 thus producing many scientific reports across a broad scope of topics regarding the marine environment. Many of this data can now be found in the form of physical manuscripts, and in some cases, .pdfs of the physical copies. As such, this information is not readily accessible for data dissemination and as such needs to be digitalized and/or digitized.

For these manuscripts to serve the greatest use to researchers, the biodiversity information is to be extrapolated and then entered into a spreadsheet. This will be entered as verbatimIdentification. From here, to meet the Darwin Core Standard, the data is to be checked for up-to-date taxonomic classifications to ensure that accurate scientific names are represented. This information, when retrieved is then to entered under the DwC term scientificName.

GBIF'S SPECIES NAME MATCHING TOOL

A streamlined process has been created to allow for easily retrievable taxon information. This can be done in one of two ways:

1) Through GBIF's Species Name Matching Tool
2) The R script, available here 

For this process to work, a compiled list of the verbatimIdentification is to be created from the literature and renamed under "scientificName" under a new .csv file. This must be done to ensure that the GBIF tool can read the file and generate the taxonomic information.

Once on the webpage, the .csv file is to be uploaded and it will output a page that shows the taxonomic classifications of the species. On this page, there will be options to resolve any doubtful classes or perceived errors in the species' names given. Once all resolutions have been made, the final .csv file can be downloaded by clicking "Generate CSV" at the bottom right of the page.

R SCRIPT

An R script was created that is able to read the .csv file with the verbatimIdentification and output a new .csv file with the taxonomic break downs.

It is important to note that for the R script to properly interact with the GBIF API, the .Renviron file must be edited to reflect a GBIF account and passcode. For IMA's general purposes, the following code is to be copied and pasted into the .Renviron file and saved:

```
PATH="${RTOOLS40_HOME}\usr\bin;${PATH}"

GBIF_USER="ima_gbif"

GBIF_PWD="IMAGBIF*1"

GBIF_EMAIL=gbifadmin@ima.gov.tt
```

To run the actual [1]R script, copy and paste the following script into R Studio ensuring to properly label the file pathway. It is important to note that R Studio cannot read backward slashes so all slashes must be forward i.e. " / ".

#These commands read the following libraries into RStudio to allow the necessary function commands to run and interact with GBIF's API.

```
library(rgbif)

library(plyr)

library(readr)
```

#This line reads the .csv file containing the scientificnames to be double checked.

```
mynames1 <- read_csv (" ~ insert pathway to .csv file containing species information~ ")
```

#This line initializes the output file ensuring that no information is already contained within it.

```
mygbif <- NULL
```

#This For Loop reads each scientific name in the file, prints it as verbatimName, and then binds the taxonomic backbone information to the relevant names.

[1] The R script generates many terms that are NOT DwC terms. Therefore, it is important to note which of the columns produced can be used for publication onto GBIF's IPT.

```
for (i in 1:nrow(mynames1)){
  #for (i in 1:10){
  print(mynames1$scientificName[i])
  verbatimName <- mynames1$scientificName[i]
  g1 <- name_backbone(mynames1$scientificName[i])
  g1 <- cbind(verbatimName,g1)
  mygbif <- rbind.fill(mygbif,g1)
}
#This line creates an output .csv file with the relevant information.
write.csv(mygbif," ~ insert desired name of output file~ .csv",row.names = F)
```

HOW TO ENTER DATA INTO THE APPROPRIATE TEMPLATES

The IMA data is extensive and either in the form of reports or raw data. Both sets of data need to be analysed with the same eye. The data needs to be sorted into either:

- ☑ The event ipt Template- generally used for recording sampling events
- ☑ The occurrence ipt Template- generally used for recording species data based on locality and often includes coordinate data for mapping purposes
- ☑ The taxon core template- generally used for classification of specimens based on taxa

By clicking on the above links, the appropriate templates, as updated by GBIF, will appear along with more information on each template. Simply download the templates to begin the data entry process.

It is important to note that additional data can be added should an extension core exist that allows it.

There is data at the IMA that goes beyond biodiversity information. If there is, however, some biodiversity information, then a checklist (taxon core) can be used to at least reference the text and associated biodiversity data so that the data can be posted to GBIF. Additional details can be added to allow the users of the data to gain a better understanding of the project as well as let them know what other non-biodiversity data s found within the dataset.

Integrated Publishing Toolkit

Chapter

4

Data that has already been standardized to meet Darwin Core Standards can be uploaded to GBIF's website via their Integrated Publishing Toolkit (IPT). This site has two modes, namely the test mode and production mode:

(I) Test Mode

(II) Latin America and Caribbean (LAC) IPT

The details steps for publishing data are available in the IPT User Manual. The link is show below.

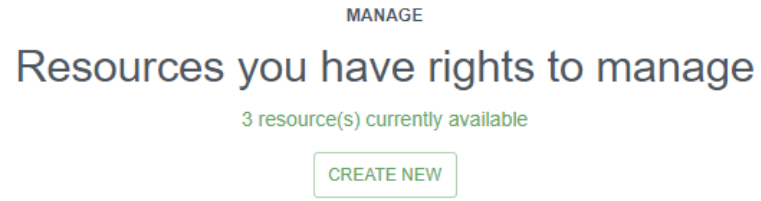
- How to publish biodiversity data through GBIF.org :: GBIF IPT User Manual

For this process, an administrative email and password is required. GBIF Secretariat Helpdesk at helpdesk@gbif.org can be contacted to access the IPT.

A general idea of publishing data is shown below.

1. Login into Test Mode. Then click on “Manage Resources” in the Top Right Corner

2. This will take you to a secondary page which showcases all the resources that you have rights to edit and also allows you to “Create New”.



3. When you click “Create New”, a page will load that allows you to manage the new resource and asks you to create a shortlist name. This name is then used in the url once the resource is created.

4. Once the resource has been created, an overview page will be generated that allows the user to add:

    a. The source data: This is the .csv file containing the resource data (either the taxon core, sampling event or occurrence data)
    b. Darwin Core Mappings: this simply asks the user to define the data set type (taxon core vs. sampling event

vs. occurrence). Further mappings can be added if the data contains extra spreadsheets- such as reference citations- or images.

c. Metadata: this is where the user fills in information regarding what the data entails, the years it was collected over, the spatial distribution of the data, project information, etc.

5. Once the metadata and source data has been added and mapped, the data should then be reviewed by GBIF Secretariat Helpdesk at helpdesk@gbif.org for a final review before being made public.
6. Make the dataset PUBLIC on the 'Visibility' section in the overview of the dataset.
7. Go to 'Published versions' and push the button PUBLISH; after this, the dataset will be published on the IPT.
8. Finally, go to the 'Visibility' section again and push REGISTER to index the dataset in the GBIF-UAT data portal (test environment).

THE DATA IS NOW PUBLIC AND UPLOADED ONTO GBIF!

All data published on GBIF by IMA can be seen on GBIF IMA webpage. The data can be edited from the LAC Portal previpously mentioned, ensuring an accurate version history and description of changes is provided. Major changes to the source data can result in a new url being made for the data, which should be reflected in all reports that need to link the data to GBIF.

References

Darwin Core quick reference guide. Biodiversity Information Standards (TDWG). https://dwc.tdwg.org/terms/

GBIF Secretariat (2021) GBIF Biodiversity Data Mobilization Course. 12th edition. GBIF Secretariat: Copenhagen. https://doi.org/10.35035/ce-c6cr-6w42.

GBIF (2021) Darwin Core Archives – How-to Guide, version 2.2. Copenhagen: GBIF Secretariat. https://ipt.gbif.org/manual/en/ipt/2.5/dwca-guide

GBIF (2023) GBIF Integrated Publishing Toolkit (IPT) User Manual Version 2.5 (Aug 2021) - 2.7.3 (March 2023) . https://ipt.gbif.org/manual/en/ipt/latest/