INSTITUTE OF MARINE AFFAIRS

DARWIN CORE STANDARD BEST PRACTICES

Marine Biodiversity Data Hub Manual

INSTITUTE OF MARINE AFFAIRSS

Marine Biodiversity Data Hub Manual

Hilltop Lane, Chaguaramas
Trinidad & Tobago

Prepared by: Cherisse Persad
Associate Professional
Biodiversity and Ecology Research Programme

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)

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 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.

Chapter 2

The Marine Reference Collection

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:

(I) The event_ipt_Template- generally used for recording sampling events

(II) The occurrence_ipt_Template- generally used for recording species data based on locality and often includes coordinate data for mapping purposes

(III) 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 main cores that accommodate the data collected and showcased in the MRC include:

- occurrenceID
- basisofRecord
- catalogNumber
- type
- othercatalogNumber
- individualCount
- sex
- lifeStage
- year
- month
- day
- eventDate
- eventRemarks
- verbatimDepth
- minimumDepthinMeters
- maximumDepthinMeters
- habitat
- parentEventID
- preparations
- verbatimIdentification
- scientificName
- Taxa (not the DwC term, the terms listed below are accepted)
    - kingdom
    - phylum
    - class
    - order
    - family
    - genus
- taxonomicStatus
- recordedBy
- verbatimLocality
- locality
- country
- countryCode
- verbatimCoordinates
- decimalLatitude
- decimalLongitude
- coordinateUncertaintyInMetres
- verbatimCoordinateSystem
- geodeticdatum
- georeferenceProtocol
- georeferenceSources
- georeferenceVerificationStatus
- georeferencedBy
- georeferencedDate
- georeferencedProtocol
- georeferenceSource
- georeferenceVerificationStatus
- occurrenceRemarks

📖 occurrenceID

- Refers to a globally unique identifier that is different from the digital record of the occurrence

- In the event of an occurrence not having a unique identifier, one can be constructed based on a combinate of unique identifiers within that same record

📖 basisofRecord

- The specific nature of the data records being inputted.

- In the case of the MRC, these records are described as "PreservedSpecimen". Other acceptable examples for this field, but not specific to the MRC, are: PreservedSpecimen, FossilSpecimen, LivingSpecimen, MaterialSample, Event, HumanObservation, MachineObservation, Taxon, Occurrence, MaterialCitation

📖 catalogNumber

- An identifier (preferably unique) for the record within the data set or collection
- In the case of the MRC, this refers to the digital record number of the image of the specimen being recorded.
- This is to be updated to ensure each species in the MRC has its own unique identifier. Once the process is completed then this Manual will be updated to reflect the Standardization process determined.

📖 type

- The nature or genre of the resource
- Examples of this record include StillImage, MovingImage, Sound, PhysicalObject, Event or Text.
- In the case of the MRC, the type was recorded as "PhysicalObject" for all the specimens digitized.

📖 othercatalogNumber

- Refers to a list (concatenated and separated) of previous or alternate fully qualified catalog numbers or other human-used identifiers for the same Occurrence, whether in the current or any other data set or collection.

- For the MRC, this refers to the original identifiers previously given to the specimens. The same system was not used throughout the years, thus the MRC

📖 **individualCount**

- The number of individuals present at the time of the Occurrence.
- In the case of the MRC, this would be how many specimens of that particular species is in the collection.

📖 **sex**

- The sex of the biological individual(s) represented in the Occurrence.
- The only acceptable forms for input are male, female or hermaphrodite.

📖 **lifeStage**

- The age class or life stage of the Organism(s) at the time of the occurrence.
- Examples include Adult or Juvenile.

📖 **year**

- Refers to the year of the occurrence or sampling event. Entered as yyyy
- If missing, left blank

📖 **month**

- Refers to the month of the occurrence or sampling event. Entered as a single digit for months January through September (I.E. 1,2,3… ,9)
- If missing, left blank

📖 **day**

- Refers to the day of the occurrence or sampling event. Entered as a single digit when it is days 1 through 9 of the month. Entered in dd format for all others.
- If missing, left blank

📖 **eventDate(yyyy-mm-dd)**

- Combination of the year, month and day data entered in the format yyyy-mm-dd. This format was chosen as it one of the accepted date formats used by the IPT.

NB: For any dates that spanned multiple days, months or years, a forward slash "/" was used to indicate the date ranges. For example, a specimen collected on a voyage spanning the 23-29th of February 2022 will be written as 2022-02-23/29.

📖 **eventRemarks**

- Comments or notes about the event. This can include locality clarification remarks, conditions at the time of sampling, etc.

📖 **verbatimDepth**

- The original description of the depth below the surface. Based on the scientific projects, this can be given in a variety of units. As such, minimumDepthinMeters and maximumDepthInMeters were two cores added to allow for the standardization of the depth units.

📖 **minimumDepthinMeters**

- The minimum depth refers to the lower end of the range of depth values given. In scenarios where only one depth value was given, that value was converted to metres and also placed in the minimumDepthInMeters column.
- For conversions, the excel formula of **"=*Convert(*select cell to be converted*, "original units", "units to be changed to*")** was used. The units were selected from the excel dropdown menu.
- In the case of the conversion from fathoms to metres, manual excel conversions were done as this is an outdated unit. The conversion was entered into the excel sheet as ***"= (depth value in fathoms) * 1.8288"***. This ensures that a standard conversion value was used.

📖 **maximumDepthinMeters**

- The maximum depth refers ONLY to the higher end of the range of depth values given. Records which only had one depth value DO NOT have recorded values in this column.
- For conversions, the excel formula of **"=*Convert(*select cell to be converted*, "original units", "units to be changed to*")** was used. The units were selected from the excel dropdown menu.
- In the case of the conversion from fathoms to metres, manual excel conversions were done as this is an outdated unit. The conversion was entered into the excel sheet as ***"= (depth value in fathoms) * 1.8288"***. This ensures that a standard conversion value was used.

habitat

- A category or description based on the habitat in which the event occurred.
- Examples include “Mangrove areas, deep sea, etc”

parentEventID

- An action that occurred during the collection process
- Examples include “Trawl, image capture, etc”

preparations

- A list (concatenated and separated) of preparations and preservation methods for a specimen. Values are to be separated with a vertical bar (|)
- Examples include fossil, cast, photograph, DNA extract, skin | skull | skeleton, whole animal (ETOH) | tissue (EDTA)

verbatimIdentification

- Refers to the name identified on the specimen tag.

taxa

- This is not the main DWC term. The terms are listed below and are filled out if the information is known. The information should be double checked for accuracy using WORMS.
  - kingdom
  - phylum
  - class
  - order
  - family
  - genus

taxonomicStatus

- The status of the use of the scientificName as a label for a taxon. Requires taxonomic opinion to define the scope of a taxon. Rules of priority then are used to define the taxonomic status of the nomenclature contained in that scope, combined with the experts’ opinion. It must be linked to a specific taxonomic reference that defines the concept.
- Accepted text includes: “Invalid”, “misapplied”, “homotypic synonym”, “accepted”

📖 **recordedBy**

- Refers to the person who collected the sample. In the case of multiple collectors, a pipe " | " is to be used to separate the names.

📖 **verbatimLocality**

- Refers to the original location data that was recorded at the time of collection

📖 **locality**

- Refers to the standardized names of the locality names given. This acts to correct any misspellings or errors in the original data entry done by the researcher or technician.

📖 **country**

- Refers to the name of the country where the specimen was taken from.

📖 **countryCode**

- Protocol adapted from the ISO-3166-1-alpha-2 country code
- Using this system, the country codes most frequently used include:
  - TT- Trinidad and Tobago
  - LC- St. Lucia
  - GD- Grenada
  - VC- St. Vincent and the Grenadines
  - VE- Venezuela

📖 **decimalLatitude**

- Refers to the standardized latitude of the stated occurrence Location as determined using the Georeferencing Quick Reference Guide (Zermoglio et al. 2020, https://doi.org/10.35035/e09p-h128).
- Location data is determined by finding a standardized point and using it for all locations given with that same location name when coordinates are not provided.
- Data in this core is ONLY entered using decimal degrees. Degrees, Minutes and seconds IS NOT ACCEPTED. Using such standards will mean that maps cannot be generated for the location data.
- If the location is vague or too broad of reference point, no location data is recorded. For example, samples taken from the North Coast of Trinidad and Tobago represents too wide of an area of study and thus, coordinates are not standardized for those locations.

📖 **decimalLongitude**

- Refers to the standardized longitude of the stated occurrence Location as determined using the Georeferencing Quick Reference Guide (Zermoglio et al. 2020, https://doi.org/10.35035/e09p-h128)
- Location data is determined by finding a standardized point and using it for all locations given with that same location name when coordinates are not provided.
- Data in this core is ONLY entered using decimal degrees. Degrees, Minutes and seconds IS NOT ACCEPTED. Using such standards will mean that maps cannot be generated for the location data.
- If the location is vague or too broad of reference point, no location data is recorded. For example, samples taken from the North Coast of Trinidad and Tobago represents too wide of an area of study and thus, coordinates are not standardized for those locations.

📖 **coordinateUncertaintyInMeters**.

- The horizontal distance (in meters) from the given decimalLatitude and decimalLongitude describing the smallest circle containing the whole of the Location. Leave the value empty if the uncertainty is unknown, cannot be estimated, or is not applicable (because there are no coordinates). Zero is not a valid value for this term.

📖 **verbatimCoordinateSystem**

- The coordinate format for the decimalLatitude and decimalLongitude or the verbatimCoordinates of the Location.
- In the case of the MRC, all coordinate systems used are decimal degrees.
- Degrees, minutes, seconds IS NOT an acceptable format
- UTM can be used but decimalLatitude and decimalLongitude dwc's need to be changed to verbatimLatitude and verbatimLongitude for recording purposes.

📖 **geodeticdatum**

- Refers to the Projection of the data. This is always WGS84.

📖 **georeferencedBy**

- Refers to the individual who did the georeferencing of the occurrence locations

📖 **georeferencedDate**

- Refers to the date that the occurrence location was standardized, the coordinates determined and the uncertainty measurements were taken or the last updated date.

📖 **georeferenceProtocol**

- A description or reference to the methods used to determine the spatial footprint, coordinates, and uncertainties.
- In the case of the MRC, the Protocol used is: Georeferencing Quick Reference Guide (Zermoglio et al. 2020, https://doi.org/10.35035/e09p-h128)

📖 **georeferenceSources**

- A list (concatenated and separated) of maps, gazetteers, or other resources used to georeference the Location, described specifically enough to allow anyone in the future to use the same resources.
- In the case of the MRC, the source was Maxar Technologies on Google Earth

📖 **georeferenceVerificationStatus**

- A categorical description of the extent to which the georeference has been verified to represent the best possible spatial description for the Location of the Occurrence.
- In the case of the MRC, the data would be "verified by the data custodian." Other options include "unable to georeference", "requires georeference", "requires verification" and "verified by researcher".

📖 **associatedReferences**

- A list (concatenated and separated) of identifiers (publication, bibliographic reference, global unique identifier, URI) of literature associated with the Occurrence.
- In the case of the MRC, this reflects the associated publications/scientific reports that were outputted from the occurrence that led to the specimen in question.

📖 **occurrenceRemarks**

- Comments or notes about the Occurrence.
- In the case of the MRC, this data can include remarks on the state of the specimen in collection (whether it is intact)

To access the MRC:

1) Navigate to the following link: Spreadsheet . The spreadsheet that has been standardized to GBIF Standards is the "IMA MRC.csv" file.

Formatting of the file:

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

Historical Data Sets

Chapter 3

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.

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) R script

GBIF'S SPECIES NAME MATCHING TOOL

For this process to work, a compiled list of the verbatimIdentification is to be created. This must first be saved as a .csv file before it can be read by the tool. 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 that will show all classifications as well as the taxonomicStatus.

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. This can be done by copying 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. " / ".

```
library(rgbif)

library(plyr)

library(readr)

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



mygbif <- NULL

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)

}

write.csv(mygbif," ~ insert desired name of output file~ .csv",row.names = F)
```

References

Darwin Core Maintenance Group. 2021. 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.