



Structuring and Managing Biological Data



Anton Van de Putte
Yi Ming Gan
Aäron Roex



The SCAR Antarctic Biodiversity Portal



Topical collection: “Antarctic and Southern Ocean biodiversity”

- Still open for submissions
- Type of papers
 - Data Paper (Biosciences)Methods
 - OMICS Data Paper
 - R Package
 - Software Description



Why publish data?

- In order to promote international cooperation in scientific investigation in Antarctica,.., the Contracting Parties agree that, to the greatest extent feasible and practicable: scientific observations and results from Antarctica shall be exchanged and made freely available.
- Open Science

What is Metadata?

“Data about data”

How, when, why

No (individual) measurements

E.g. license, institution, sampling
method, ...

FAIR principles

Findable

Accessible

Interoperable

Reusable

- Metadata
- Infrastructures
- Standardisation
- Machine readable
- Trusted repositories



MAKE YOUR DATA FAIR



Apply a suitable usage license



Provide end users with information on
“intended use”



Make the data “as open as possible, as closed
as necessary”



Ensure that metadata remain available even if
the data cannot be accessed any more



FAIRness needs to be applied where it makes
sense

As open as possible,
as closed as needed

- Data
- (Partly) hide data
- Never hide metadata
- E.g. GDPR, Threatened species

Standardised data

“Data Scientists spend up to 80% of the time on data cleaning and 20 percent of their time on actual data analysis” (*Dasu and Johnson 2003*)

The first person to benefit from clean structured data is future you

Tidy Data

- For downstream operations (QC)
- Necessary for repos
- Machine readability
- Human readability

country	year	cases	population
Afghanistan	1999	15	19987071
Afghanistan	2000	66	20095360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21766	128042583

variables

country	year	cases	population
Afghanistan	1999	15	19987071
Afghanistan	2000	66	20095360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21766	128042583

observations

country	year	cases	population
Afghanistan	1999	15	19987071
Afghanistan	2000	66	20095360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21766	128042583

values

Tidy Data

catalogNumber	basisOfRecord	type	scientificName	kingdom	phylum	class	order	family	genus
MNA13174	PreservedSpecimen	PhysicalObject	Calanoides acutus (Giesbrecht, 1902)	Animalia	Arthropoda	Hexanauplia	Calanoida	Calanidae	Calanoides
MNA13175	PreservedSpecimen	PhysicalObject	Harpacticus furcifer (Edwards H., 1840)	Animalia	Arthropoda	Hexanauplia	Harpacticoida	Harpacticidae	Harpacticus
MNA13176	PreservedSpecimen	PhysicalObject	Metridia gerlachei (Giesbrecht, 1902)	Animalia	Arthropoda	Hexanauplia	Calanoida	Metridinidae	Metridia
MNA13177	PreservedSpecimen	PhysicalObject	Metridia gerlachei (Giesbrecht, 1902)	Animalia	Arthropoda	Hexanauplia	Calanoida	Metridinidae	Metridia
MNA13178	PreservedSpecimen	PhysicalObject	Tisbe gracilipes (Scott T., 1912)	Animalia	Arthropoda	Hexanauplia	Harpacticoida	Tisbidae	Tisbe

Variable

Datapoint

Cookiecutter project structure

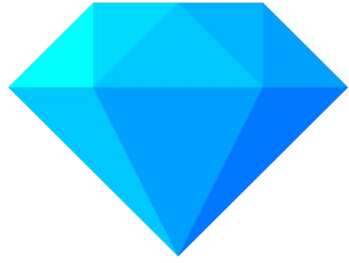
```
├── README.md          <- The top-level README.
├── data
│   ├── external      <- Data from third party sources.
│   ├── interim       <- Intermediate data that has been transformed.
│   ├── processed     <- The final, canonical data sets for modeling.
│   └── raw           <- The original, immutable data dump.
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
├── src                <- Source code for use in this project.
│   ├── data          <- Scripts to download or generate data.
│   │   └── make_dataset.R
│   ├── clean         <- Scripts to clean data.
│   │   └── clean_dataset.R
│   └── visualization <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.R
├── reports           <- Generated QC reports.
│   └── figures       <- Generated graphics and figures to be used in reporting.
```

Give good names to your files

- Machine readable
- human readable
- orderable
- use of separators
- uniform style

```
~/Desktop/projects/01_cruise-reports ▷  
2018-12-15_PS117_Cape-Town.pdf  
2019-02-09_PS118_Punta-Arenas.pdf  
2019-04-13_PS119_Punta-Arenas.pdf  
2020-06-04_PS122-4_Arctic-Ocean.pdf  
2020-08-12_PS122-5_Arctic-Ocean.pdf
```

OpenRefine



- **1. Explore Data**
- **2. Clean and Transform Data**
- **3. Reconcile and Match Data**



This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method **key collision**

Keying Function **fingerprint**

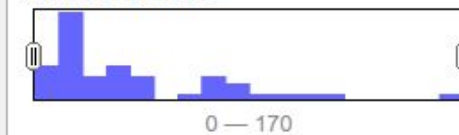
48 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	63	<ul style="list-style-type: none">Pune Vidhyapeeth Gate (33 rows)Pune Vidhyapeeth Gate (17 rows)Pune Vidhyapeeth Gate (13 rows)	<input type="checkbox"/>	Pune Vidhyapeeth Gate
3	101	<ul style="list-style-type: none">Hadapsar Gadital (83 rows)Hadapsar Gadital (14 rows)Hadapsar Gadital (4 rows)	<input type="checkbox"/>	Hadapsar Gadital
3	12	<ul style="list-style-type: none">Devachi Uruli Phata (8 rows)Devachi Uruli Phata (2 rows)Uruli Devachi Phata (2 rows)	<input type="checkbox"/>	Devachi Uruli Phata
2	2	<ul style="list-style-type: none">SRP Stadium (1 rows)SRP Stadium (1 rows)	<input type="checkbox"/>	SRP Stadium
2	8	<ul style="list-style-type: none">Khandoba Mandir Corner (6 rows)Khandoba Mandir corner (2 rows)	<input type="checkbox"/>	Khandoba Mandir Corner
2	7	<ul style="list-style-type: none">St Meera College (5 rows)ST Meera College (2 rows)	<input type="checkbox"/>	St Meera College
2	67	<ul style="list-style-type: none">Gunjan Corner (62 rows)	<input type="checkbox"/>	Gunjan Corner

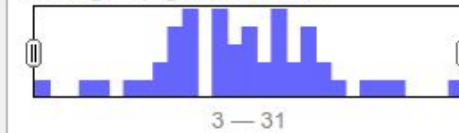
Choices in Cluster



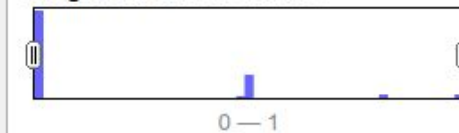
Rows in Cluster



Average Length of Choices

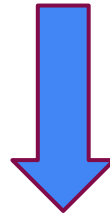


Length Variance of Choices



Machine readability: Atomize columns

verbatimCoordinates	date	taxonomy
58°28'30''W,62°08'00''S	2022-08-05	Rhincalanus gigas, Brady (1883)
62°08'10''W,61°18'20''S	2022-08-09	Rhincalanus gigas, Brady (1883)
61°58'40''W,61°48'50''S	2022-08-10	Rhincalanus gigas, Brady (1883)



decimalLatitude	decimalLongitude	year	month	day	acceptedTaxonName	scientificNameAuthorship
-62.1333	-58.4750	2022	08	05	Rhincalanus gigas	Brady,1883
-61.3055	-62.1361	2022	08	09	Rhincalanus gigas	Brady,1883
-61.8139	-61.9778	2022	08	10	Rhincalanus gigas	Brady,1883

Use Persistent identifiers

Term	Example
scientificNameID	urn:lsid:ipni.org:names:37829-1:1.3
identifiedByID	https://orcid.org/0000-0002-1825-0097
taxonID	http://species.gbif.org/abies_alba_1753
materialSampleID	06809dc5-f143-459a-be1a-6f03e63fc083

Exchanging and publishing data



Darwin Core **standard** is a set of terms

Occurrence

occurrenceID	catalogNumber	recordNumber	recordedBy	recordedByID
individualCount	organismQuantity	organismQuantityType	sex	lifeStage
reproductiveCondition	behavior	establishmentMeans	degreeOfEstablishment	
pathway	georeferenceVerificationStatus	occurrenceStatus	preparations	
disposition	associatedMedia	associatedOccurrences	associatedReferences	
associatedSequences	associatedTaxa	otherCatalogNumbers	occurrenceRemarks	

Occurrence		Class
Identifier	http://rs.tdwg.org/dwc/terms/Occurrence	
Definition	An existence of an Organism (sensu http://rs.tdwg.org/dwc/terms/Organism) at a particular place at a particular time.	
Comments		
Examples	A wolf pack on the shore of Kluane Lake in 1988. A virus in a plant leaf in the New York Botanical Garden at 15:29 on 2014-10-23. A fungus in Central Park in the summer of 1929.	

occurrenceID		Property
Identifier	http://rs.tdwg.org/dwc/terms/occurrenceID	
Definition	An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.	
Comments	Recommended best practice is to use a persistent, globally unique identifier.	
Examples	http://arctos.database.museum/guid/MSB:Mamm:233627 , 000866d2-c177-4648-a200-ead4007051b9 , urn:catalog:UWBM:Bird:89776	

Record-level

Occurrence

Organism

MaterialSample

Event

Location

GeologicalContext

Identification

Taxon

MeasurementOrFact

ResourceRelationship

UseWithIRI

LivingSpecimen

PreservedSpecimen

FossilSpecimen

MaterialCitation

HumanObservation

MachineObservation

A data package is needed for data exchange

Occurrence

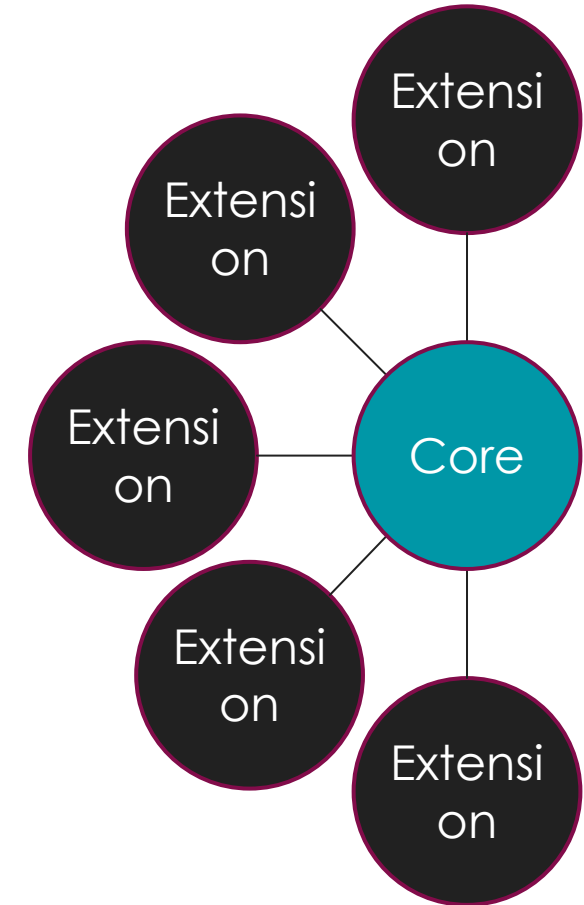
occurrenceID	catalogNumber	recordNumber	recordedBy	recordedByID
individualCount	organismQuantity	organismQuantityType	sex	lifeStage
reproductiveCondition	behavior	establishmentMeans	degreeOfEstablishment	
pathway	georeferenceVerificationStatus	occurrenceStatus	preparations	
disposition	associatedMedia	associatedOccurrences	associatedReferences	
associatedSequences	associatedTaxa	otherCatalogNumbers	occurrenceRemarks	

Occurrence	Class
Identifier	http://rs.tdwg.org/dwc/terms/Occurrence
Definition	An existence of an Organism (sensu http://rs.tdwg.org/dwc/terms/Organism) at a particular place at a particular time.
Comments	
Examples	A wolf pack on the shore of Kluane Lake in 1988. A virus in a plant leaf in the New York Botanical Garden at 15:29 on 2014-10-23. A fungus in Central Park in the summer of 1929.

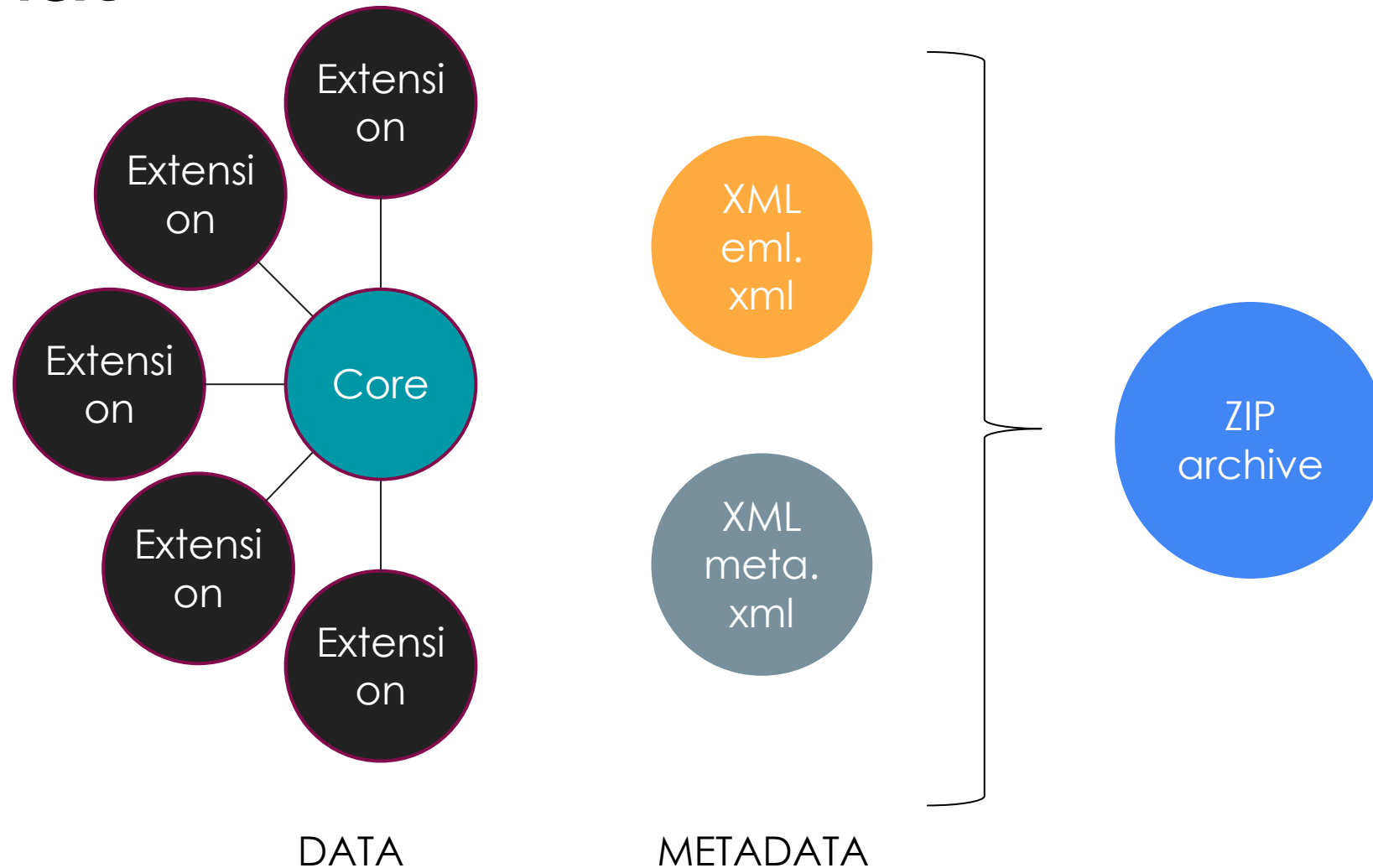
occurrenceID	Property
Identifier	http://rs.tdwg.org/dwc/terms/occurrenceID
Definition	An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.
Comments	Recommended best practice is to use a persistent, globally unique identifier.
Examples	http://arctos.database.museum/guid/MSB:Mamm:233627 , 000866d2-c177-4648-a200-ead4007051b9 , urn:catalog:UWBM:Bird:89776

- Record-level
 - Occurrence**
 - Organism
 - MaterialSample
 - Event
 - Location
 - GeologicalContext
 - Identification
 - Taxon
 - MeasurementOrFact
 - ResourceRelationship
 - UseWithIRI
-
- LivingSpecimen
 - PreservedSpecimen
 - FossilSpecimen
 - MaterialCitation
 - HumanObservation
 - MachineObservation

subset of terms



Darwin Core **archive** is a data exchange format



Core: main observation

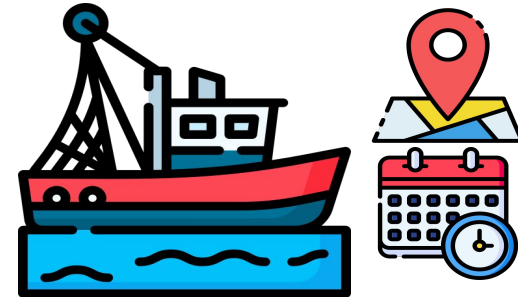
Occurrence Core

- Organisms at specific time and place
- Required terms:
manual.obis.org/darwin_core.html
- E.g. Whales spotted



Event Core

- Sampling events
- Occurrence extension
- Required terms:
manual.obis.org/darwin_core.html
- E.g. Trawls, sampling station



Extensions: other data



34 registered extensions
<https://rs.gbif.org/extensions.html>



Data not in Core



Needs pointer to Core



E.g. Temperature, Pictures, ...

Ecological metadata Language (Eml.xml)

Automatically generated

Basic Metadata

Geographic Coverage

Taxonomic Coverage

Temporal Coverage

Keywords

Associated Parties

Project Data

Sampling Methods

Citations

Collection Data

meta.xml



RELATIONSHIP
BETWEEN FILES



AUTOMATICALLY
GENERATED



HUMAN AND
MACHINE READABLE

```
<?xml version="1.0"?>
<eml:eml
  packageId="eml.1.1" system="knb"
  xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:stmml="http://www.xml-cml.org/schema/stmml-1.1"
  xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1
eml.xsd">

<dataset>
  <title>Data from Cedar Creek LTER on productivity and species
richness
  </title>
  <creator id="clarence.lehman">
    <individualName>
      <salutation>Mr.</salutation>
      <givenName>Clarence</givenName>
      <surName>Lehman</surName>
    </individualName>
  </creator>
  <contact>
    <references>clarence.lehman</references>
  </contact>
</dataset>
</eml:eml
```

```
<archive xmlns="http://rs.tdwg.org/dwc/text/" metadata="eml.xml">

  <core encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n"
fieldsEnclosedBy="" ignoreHeaderLines="1"
rowType="http://rs.tdwg.org/dwc/terms/Occurrence">
    <files>
      <location>occurrence.txt</location>
    </files>
    <id index="0"/>
    <field index="1" term="http://purl.org/dc/terms/license"/>
    <field index="2" term="http://purl.org/dc/terms/bibliographicCitation"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/institutionCode"/>
    <field index="4" term="http://rs.tdwg.org/dwc/terms/collectionCode"/>
    <field index="5" term="http://rs.tdwg.org/dwc/terms/basisOfRecord"/>
    <field index="6" term="http://rs.tdwg.org/dwc/terms/dynamicProperties"/>
    <field index="7" term="http://rs.tdwg.org/dwc/terms/occurrenceID"/>
    <field index="8" term="http://rs.tdwg.org/dwc/terms/eventDate"/>
    <field index="9" term="http://rs.tdwg.org/dwc/terms/country"/>
    <field index="10"
term="http://rs.tdwg.org/dwc/terms/minimumDepthInMeters"/>
    <field index="11" term="http://rs.tdwg.org/dwc/terms/decimalLatitude"/>
    <field index="12" term="http://rs.tdwg.org/dwc/terms/decimalLongitude"/>
    <field index="13" term="http://rs.tdwg.org/dwc/terms/scientificNameID"/>
    <field index="14" term="http://rs.tdwg.org/dwc/terms/scientificName"/>
    <field index="15" term="http://rs.tdwg.org/dwc/terms/kingdom"/>
    <field index="16" term="http://rs.tdwg.org/dwc/terms/phylum"/>
    <field index="17" term="http://rs.tdwg.org/dwc/terms/class"/>
    <field index="18" term="http://rs.tdwg.org/dwc/terms/order"/>
    <field index="19" term="http://rs.tdwg.org/dwc/terms/family"/>
    <field index="20" term="http://rs.tdwg.org/dwc/terms/genus"/>
  </core>
</archive>
```



Darwin Core quick reference guide

- <https://dwc.tdwg.org/terms/>

Darwin Core extensions

- <https://rs.gbif.org/extension/>

Darwin core validator

- <https://www.gbif.org/tools/data-validator/about>

Darwin core instructions

- <https://manual.obis.org/>

MAKE YOUR DATA FAIR



Make a plan for the data before you start a project!



Collect detailed descriptive information (= metadata) throughout



Use standards and formats common to your discipline



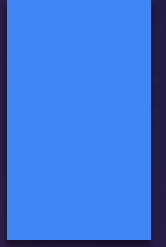
Store the data in a trusted & sustainable repository or data center



See to that the data gets persistent identifiers (DOIs)



Recording and publishing data



Which variables? TDWG Vocabulary

Some Required by repo:

Many more optional or recommended

Occurrence Core	Event Core
occurrenceID	eventID
basisOfRecord	
scientificName	
scientificNameID	
occurrenceStatus	
decimalLongitude	
decimalLatitude	
eventDate	eventDate

Occurrence Core	Event Core
occurrenceStatus	parentEventID
recordedBy	minimumDepthInMeters
taxonRank	
kingdom	
coordinateUncertaintyInMeters	
individualCount	

Useful Links

- <https://manual.obis.org/index.html>
- <https://dwc.tdwg.org/terms/>

Consistency

Geographic coordinates

- Type of decimal point
- Order of longitude, latitude
- Degree minute second vs degree decimal minute
- Padding with 0

verbatimCoordinates

58°28'30"W, 62°7'0,10"S

58°28'42,00"W, 62°7'15,00"S

62°08.0958'S, 058°24,1625'W

62°10.4680'S, 058°25.2137'W

62°09.7174'S, 058°21,5886'W

62°11.2075'S, 058°18,9951W

62°11.9764'S, 058°22.5800W

62°12.1942'S, 058°23.4483'W

62°53.6407'S, 58°26.8232'W

Taxonomy

Match scientific name to taxon using WoRMS taxon match service

The screenshot displays the WoRMS (World Register of Marine Species) website. At the top, the logo 'WoRMS World Register of Marine Species' is visible. A navigation bar includes links for Home, About, Subregisters, Users, Photogallery, Documents, LifeWatch, and Contribute. Below this is a search bar and a secondary navigation bar with icons for Taxa, Literature, Distribution, Specimen, Editors, Statistics, Tools, Manual, and Log in.

The main content area features a banner with the text 'An authoritative classification and catalogue of marine names' and the WoRMS logo. Below the banner are three images: a black crustacean, a red crustacean, and a microscopic organism.

On the right side, there is a 'Latest taxon additions' section, updated on 2020-10-27T18:27. It lists several species with their dates of addition:

- Mitra porcata Reeve, 1844 (Added: 2020-10-27)
- Trophon muricatus Hinds, 1844 (Added: 2020-10-27)
- Erylus mastoideus (Schmidt, 1880) (Added: 2020-10-27)
- Ancorina individuosa (Added: 2020-10-27)
- Ecionema rotundum (Added: 2020-10-27)

Below the banner, there is a 'News' section. The first article is titled 'The World Flora Online & the International Compositae Alliance join forces', dated 2020-10-26 10:58:19 by Vandepitte, Leen. The text states: 'About six months after the official launch of the Global Compositae Database through the Aphia infrastructure, its content is now being shared with the World Flora Online. The collaboration between these two initiatives has become official through a Memorandum of Understanding (MoU). ...'. A 'Read more' link is provided.


At the bottom right, there is a 'Tweets by @WRMarineSpecies' section. It shows a tweet from LifeWatch VLIZ (@LifeWatchVLIZ) stating: 'This week, @WRMarineSpecies and @LifeWatchVLIZ are present in the virtual sDevTrait workshop, organized by @idiv'.

Taxonomy

Keep original scientific name

Use persistent identifier to refer to the taxon in a checklist

Do not change to accepted scientific name



scientificName	scientificNameAuthorship	scientificNameID	taxonomicStatus	taxonRank
Akanthophoreus antarcticus	(Vanhöffen, 1914)	urn:lsid: marinespecies.org :taxname:448316	unaccepted	species
Akanthophoreus multiserratus	(Hansen, 1913)	urn:lsid: marinespecies.org :taxname:136343	unaccepted	species



Persistent identifiers

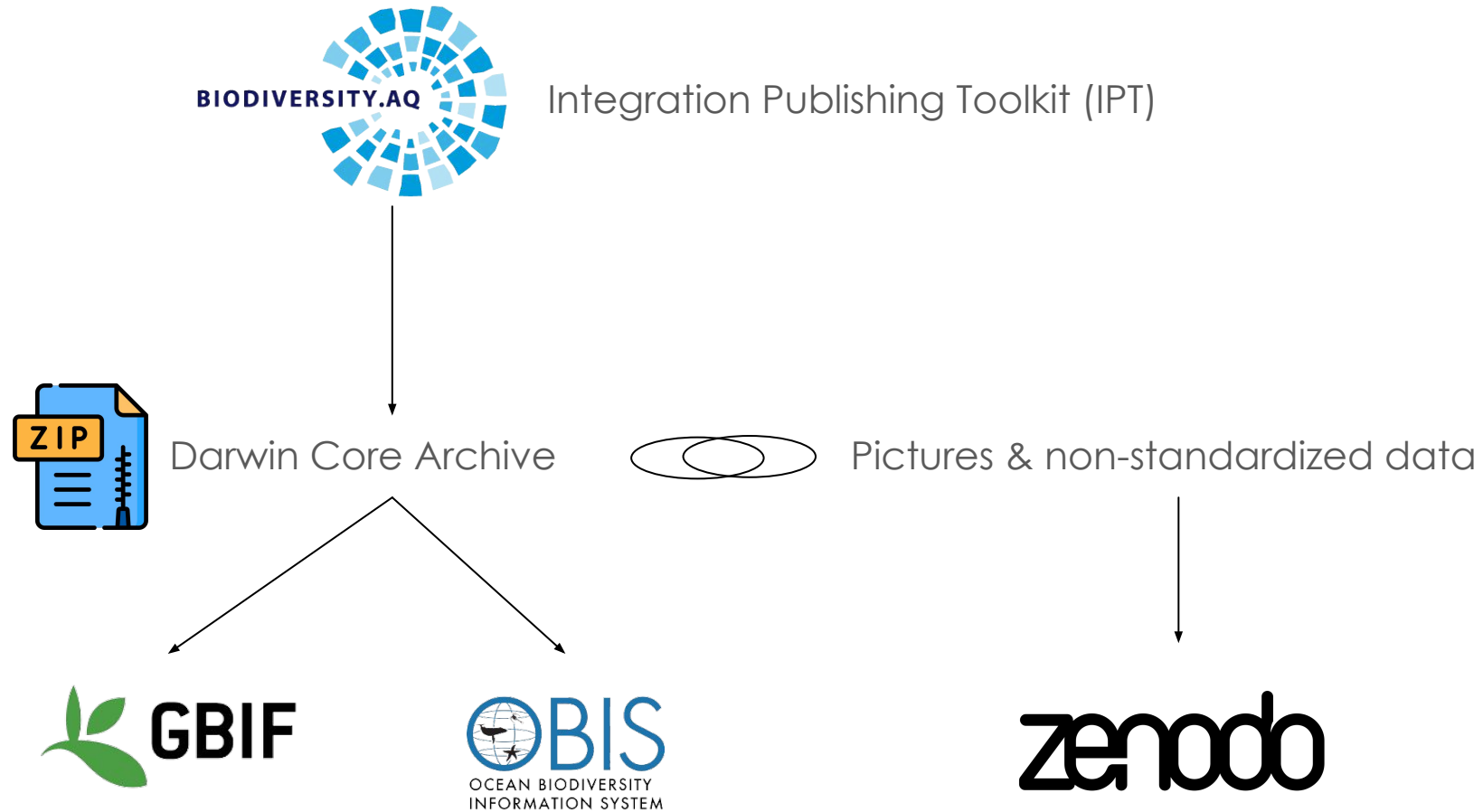
Useful links

- <https://dwc.tdwg.org/terms/>
- <https://www.marinespecies.org/>
- <https://www.doi.org/>



Uploading to
repository

Biodiversity repositories





[Home](#)

[Manage Resources](#)

[Administration](#)

[About](#)

Resource Title [Polychaetes from the JR17003 Expedition](#)

Basic Metadata

Please enter all the mandatory properties on the Basic Metadata page, and then continue entering metadata in the other pages that are applicable to your resource. The more metadata you provide, the greater the chance that your resource will be found, reused by other researchers, and cited.

Title*

Polychaetes from the JR17003 Expedition

Publishing Organisation*

[i](#) Select an organisation [v](#)

Type*

[i](#) Occurrence [v](#)

Metadata Language*

[i](#) English [v](#)

Update Frequency*

[i](#) Unknown [v](#)

Subtype

[i](#) Select a subtype [v](#)

Data Language*

[i](#) English [v](#)

Data Licence*

[i](#) No licence selected [v](#)

Section

[Basic Metadata](#)

[Geographic Coverage](#)

[Taxonomic Coverage](#)

[Temporal Coverage](#)

[Keywords](#)

[Associated Parties](#)

[Project Data](#)

[Sampling Methods](#)

[Citations](#)

[Collection Data](#)

[External links](#)

[Additional Metadata](#)

Darwin Core Archive Dataset metadata is **very similar** to Pensoft Data Paper



not in Darwin Core Archive dataset metadata

Screenshot from ARPHA Writing Tool

Links

<https://GBIF.org>

<https://OBIS.org>

www.biodiversity.aq

<https://ipt.biodiversity.aq/>

<https://zenodo.org/>

Are your data FAIR?

- ANDS FAIR self-assessment tool

<https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/>

- FAIRDAT assessment tool (prototype)

<https://www.surveymonkey.com/r/fairdat>

- How FAIR is your data?

<https://forms.gle/eBagszpWKVz5NKpp7>



How to use ARPHA writing tool

arpha Writing Tool (AWT) Feedback Tips and Tricks Ms Yi Ming Gan

Collaborative authoring, reviewing and publishing in one place, for the first time!

Start manuscript

Convert from file

My recent manuscripts



- Application of Humboldt Extension to Real-world Cases
- Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and...

View dashboard

Select BDJ & Data Paper (Biosciences) article type

To start a manuscript select a journal and an article type

-  Biodiversity Data Journal
-  Research Ideas and Outcomes
-  One Ecosystem
-  BioDiscovery
-  Biodiversity Information Science and Standards
-  Food and Ecological Systems Modelling Journal
-  ARPHA Conference Abstracts
-  Viticulture Data Journal
-  Biosystematics and Ecology

- | Research ideas | Early research outcomes | Brief research outcomes |
|--|---|--|
| <input type="radio"/> Data Management Plan | <input type="radio"/> Applied Study | <input type="radio"/> Commentary  |
| <input type="radio"/> Grant Proposal | <input type="radio"/> Case Study | <input type="radio"/> Conference Abstract |
| <input type="radio"/> PhD Project Plan | <input type="radio"/> Clinical Case Studies | <input type="radio"/> Correspondence |
| <input type="radio"/> PostDoc Project Plan | <input type="radio"/> Data analytics | <input type="radio"/> Ecosystem Accounting Table |
| <input type="radio"/> Research Idea | <input checked="" type="radio"/> Data Paper (Biosciences)  | <input type="radio"/> Ecosystem Inventory |

ARPHA writing tool interface

The screenshot displays the ARPHA writing tool interface. At the top, a navigation bar includes links for 'View dashboard', 'Collections', 'Email contributors', 'Helpdesk', 'Tips and tricks', 'Tutorial', and 'Revision history'. Below this, a header bar shows 'Data Paper (Biosciences)' on the left, a central warning 'This is a read-only version!' with a red triangle icon, and 'Comments', 'Resolved', and 'Filter' on the right.

The main content area is divided into a left sidebar and a central panel. The sidebar contains a list of navigation options: 'Authors', 'Contributors', 'Article metadata' (with sub-items: 'Title', 'Abstract & Keywords', 'Classifications', 'Funder'), 'Introduction', 'General description', 'Project description', 'Sampling methods', 'Geographic coverage', 'Taxonomic coverage', 'Traits coverage' (with sub-item: 'Data coverage of traits'), 'Temporal coverage', 'Collection data', 'Usage licence', 'Data resources', 'Additional information', 'Acknowledgements', 'Author contributions', 'References', 'Supplementary files', and 'Figures'.

The central panel displays the article details for 'Biodiversity Data Journal : Data Paper (Biosciences)'. The title is 'Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and Weddell Sea - data'. The authors listed are Jamie Maxwell[‡], Yi Ming Gan[§], Claudia Arango[|], Jana S Doemel[¶], A. Louise Allcock[‡], Anton P. van de Putte[§], and Huw Griffiths[#]. Below the authors, their affiliations are provided: [‡] National University of Ireland, Galway, Galway, Ireland; [§] Royal Belgian Institute of Natural Sciences, Brussels, Belgium; [|] Queensland Museum, Brisbane, Australia; [¶] University of Duisburg-Essen, Essen, Germany; and [#] British Antarctic Survey, Cambridge, United Kingdom.

A box below the affiliations contains the following information: Corresponding author: Jamie Maxwell (j.maxwell4@nuigalway.ie); Academic editor: Heather Lynch; Received: 14 Dec 2021 | Accepted: 29 Apr 2022 | Published: 14 Jun 2022; © 2022 Jamie Maxwell, Yi Ming Gan, Claudia Arango, Jana Doemel, A. Louise Allcock, Anton P. van de Putte, Huw Griffiths; This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. An 'OPEN ACCESS' logo is also present. The citation is: Maxwell J, Gan YM, Arango C, Doemel JS, Allcock AL, van de Putte AP, Griffiths H (2022) Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and Weddell Sea - data. Biodiversity Data Journal 10: e79353. The DOI is <https://doi.org/10.3897/BDJ.10.e79353>.

The 'Abstract' section is titled 'Background' and contains the text: 'This dataset contains information on specimens of Southern Ocean Pycnogonida (Arthropoda), that were collected from ten different research cruises, spanning 13 years. The individual aims and objectives of each cruise can be found in their cruise reports. The specimens have been collated into a single dataset, forming the basis of J. Maxwell's PhD. The dataset will be used to investigate the community structure of Antarctic pycnogonids and the factors which influence its composition. This dataset is published by SCAR-AntOBIS under the licence CC-BY 4.0. Please follow the guidelines from the SCAR and IPY Data Policies'.

Example data paper: <https://bit.ly/3JyNfl4>

FAQ

Kudos to everyone who sent
in the example dataset

What is expected for this topical collection?

2 publications are expected, but they are almost identical



Navigation bar with links: View dashboard, Collections, Email contributors, Helpdesk, Tips and tricks, Tutorial, Revision history. A pink banner indicates "This is a read-only version!". On the right, there are links for Comments, Resolved, and Filter.

- Authors
- Contributors
- Article metadata
 - Title
 - Abstract & Keywords
 - Classifications
 - Funder
- Introduction
- General description
- Project description
- Sampling methods
- Geographic coverage
- Taxonomic coverage
- Traits coverage
 - Data coverage of traits
- Temporal coverage
- Collection data
- Usage licence
- Data resources
- Additional information
- Acknowledgements
- Author contributions
- References
- Supplementary files
- Figures

Biodiversity Data Journal : Data Paper (Biosciences) PDF

Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and Weddell Sea - data

Jamie Maxwell[‡], Yi Ming Gan[§], Claudia Arango[‡], Jana S Doemel[‡], A. Louise Alcock[‡], Anton P. van de Putte[§], Huw Griffiths[#]

[‡] National University of Ireland, Galway, Galway, Ireland
[§] Royal Belgian Institute of Natural Sciences, Brussels, Belgium
| Queensland Museum, Brisbane, Australia
[#] University of Duisburg-Essen, Essen, Germany
[#] British Antarctic Survey, Cambridge, United Kingdom

Corresponding author: Jamie Maxwell (j.maxwell4@nuigalway.ie)

Academic editor: Heather Lynch

Received: 14 Dec 2021 | Accepted: 29 Apr 2022 | Published: 14 Jun 2022

© 2022 Jamie Maxwell, Yi Ming Gan, Claudia Arango, Jana Doemel, A. Louise Alcock, Anton P. van de Putte, Huw Griffiths
This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.



Citation: Maxwell J, Gan YM, Arango C, Doemel JS, Alcock AL, van de Putte AP, Griffiths H (2022) Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and Weddell Sea - data. Biodiversity Data Journal 10: e79353. <https://doi.org/10.3897/BDJ.10.e79353>

Abstract

Background

This dataset contains information on specimens of Southern Ocean Pycnogonida (Arthropoda), that were collected from ten different research cruises, spanning 13 years. The individual aims and objectives of each cruise can be found in their cruise reports. The specimens have been collated into a single dataset, forming the basis of J. Maxwell's PhD. The dataset will be used to investigate the community structure of Antarctic pycnogonids and the factors which influence its composition. This dataset is published by SCAR-AntOBIS under the licence CC-BY 4.0. Please follow the guidelines from the SCAR and IPY Data Policies

Home About Login

Sea spiders (Arthropoda, Pycnogonida) from ten recent research expeditions to the Antarctic Peninsula, Scotia Arc and Weddell Sea - data

OCCURRENCE

Latest version published by SCAR - AntOBIS on 14 December 2021

The dataset contains information on specimens of Southern Ocean Pycnogonida (Arthropoda), that were collected from 10 different research cruises, spanning 13 years. The individual aims and objectives of each cruise can be found in their cruise reports. The specimens have been collated into a single dataset, forming the basis of J Maxwell's PhD. The dataset will be used to investigate the community structure of Antarctic pycnogonids and the factors which influence its composition. This dataset is published by SCAR-AntOBIS under the license CC-BY 4.0. Please follow the guidelines from the SCAR and IPY Data Policies (<https://www.scar.org/excom-meetings/xxxi-scar-delegates-2010-buenos-aires-argentina/4563-scar-xxxi-ip04b-scar-data-policy/file/>) when using the data. If you have any questions r... [More](#)

GBIF UUID: [1e7b6980-0842-4c4a-8b14-541b95d2ed3c](https://doi.org/10.3897/BDJ.10.e79353)
Publication date: 14 December 2021
Hosted by: SCAR - AntOBIS
License: CC-BY 4.0
[How to cite](#)

- Description
- Data Records
- Downloads
- Versions
- How to cite
- Rights
- GBIF Registration
- Keywords

Description

The dataset contains information on specimens of Southern Ocean Pycnogonida (Arthropoda), that were collected from 10 different research cruises, spanning 13 years. The individual aims and objectives of each cruise can be found in their cruise reports. The specimens have been collated into a single dataset, forming the basis of J Maxwell's PhD. The dataset will be used to investigate the community structure of Antarctic pycnogonids and the factors which influence its composition. This dataset is published by SCAR-AntOBIS under the license CC-BY 4.0. Please follow the guidelines from the SCAR and IPY Data Policies (<https://www.scar.org/excom-meetings/xxxi-scar-delegates-2010-buenos-aires-argentina/4563-scar-xxxi-ip04b-scar-data-policy/file/>) when using the data. If you have any questions regarding this dataset, please don't hesitate to contact us via the contact information provided in the metadata or via data-biodiversity-aq@naturalsciences.be.

What is expected after you submitted your manuscript to BDJ?

Someone will audit your data before your data paper can go to review

Data audit based on:

<https://zookeys.pensoft.net/about#DataQualityChecklistandRecommendations>

eventDate that conforms to ISO8601 standard

https://en.wikipedia.org/wiki/ISO_8601



Adapted from: <https://www.instagram.com/punhubinsta/>

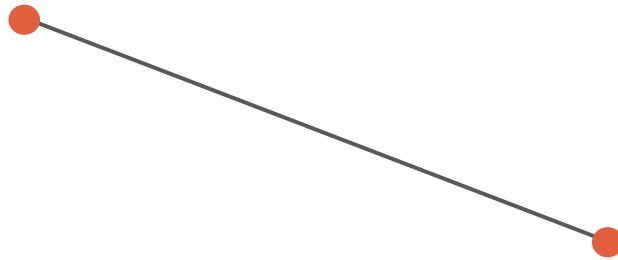
Uncertain temporal scope

All these are ok:

eventDate	year	month	day	
2022-08	2022	08		day unknown
2022	2022			month & day unknown
2020/2022				if year unknown, provide known year range

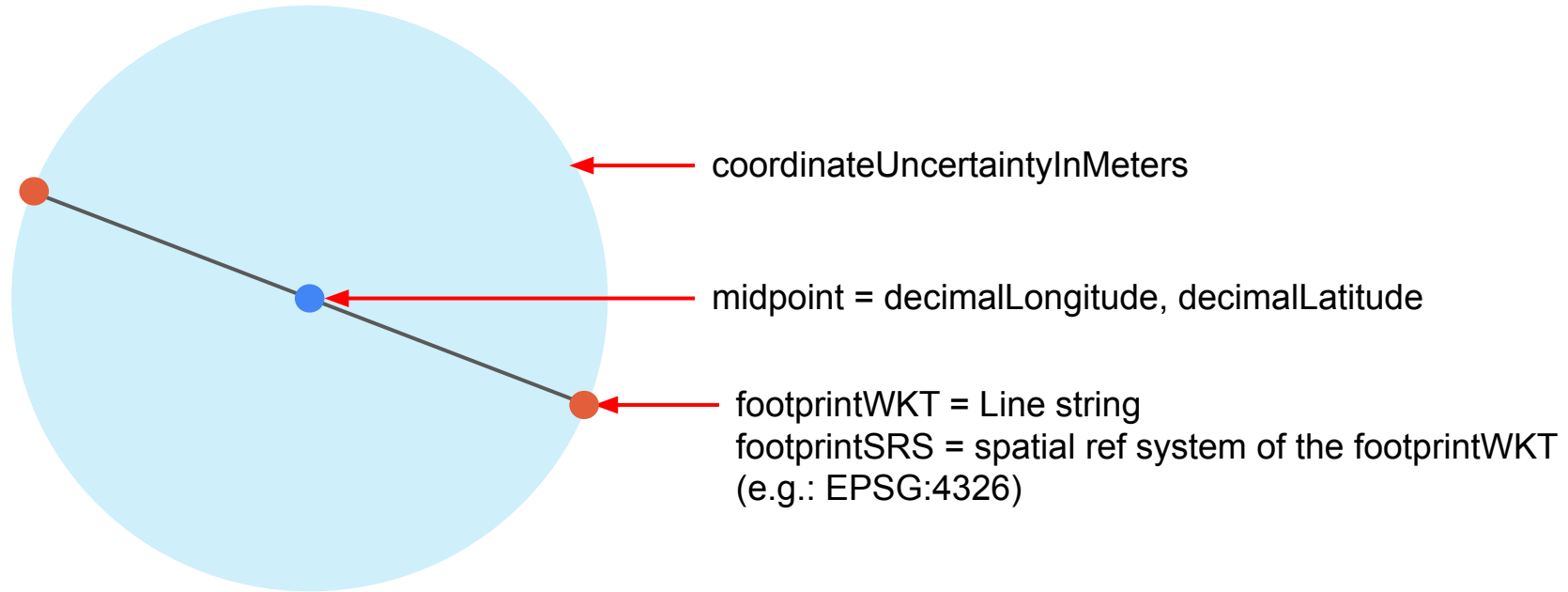
eventDate still conforms to ISO8601 standard

What if my sampling location is not a point? (e.g. trawl)



start point & end point

What if my sampling location is not a point? (e.g. trawl)



footprintWKT	footprintSRS	decimalLongitude	decimalLatitude	coordinateUncertaintyInMeters
LINestring (-29.42462 -74.67411, -29.43436 -74.67504)	EPSG:4326	-29.4295	-74.6746	153

`coordinateUncertaintyInMeters` is strongly recommended in GBIF & OBIS, please provide it regardless

Precision of decimalLongitude & decimalLatitude

decimalLongitude	decimalLatitude	geodeticDatum	coordinatePrecision	coordinateUncertaintyInMeters
-29.42	-74.67	EPSG:4326	0.01	1155
-29.4200	-74.6700	EPSG:4326	0.0001	12

Excel conversion may lead to coordinates with more decimal places than it should be

“sp.” “cf.” in identificationQualifier

scientificName	identificationQualifier	scientificNameID	genus	taxonRank
Caprella	sp.	urn:lsid:marinespecies.org:taxname:101430	Caprella	genus

Keep the most confident identification in scientificName column

Some fields are paired

organismQuantity		Property
Identifier	http://rs.tdwg.org/dwc/terms/organismQuantity	
Definition	A number or enumeration value for the quantity of organisms.	
Comments	An organismQuantity must have a corresponding organismQuantityType.	
Examples	27 (organismQuantity) with individuals (organismQuantityType). 12.5 (organismQuantity) with % biomass (organismQuantityType). r (organismQuantity) with BraunBlanquet Scale (organismQuantityType). many (organismQuantity) with individuals (organismQuantityType).	

organismQuantityType		Property
Identifier	http://rs.tdwg.org/dwc/terms/organismQuantityType	
Definition	The type of quantification system used for the quantity of organisms.	
Comments	A dwc:organismQuantityType must have a corresponding dwc:organismQuantity.	
Examples	27 (organismQuantity) with individuals (organismQuantityType). 12.5 (organismQuantity) with %biomass (organismQuantityType). r (organismQuantity) with BraunBlanquetScale (organismQuantityType).	

No vocabulary suits my data

lifeStage	
Identifier	http://rs.tdwg.org/dwc/terms/lifeStage
Definition	The age class or life stage of the Organism(s) at the time the Occurrence was recorded.
Comments	Recommended best practice is to use a controlled vocabulary .
Examples	zygote , larva , juvenile , adult , seedling , flowering , fruiting

- Use commonly used term
- include remarks in relevant ____Remarks field:
organismRemarks, eventRemarks, occurrenceRemarks etc...

No vocabulary suits my data

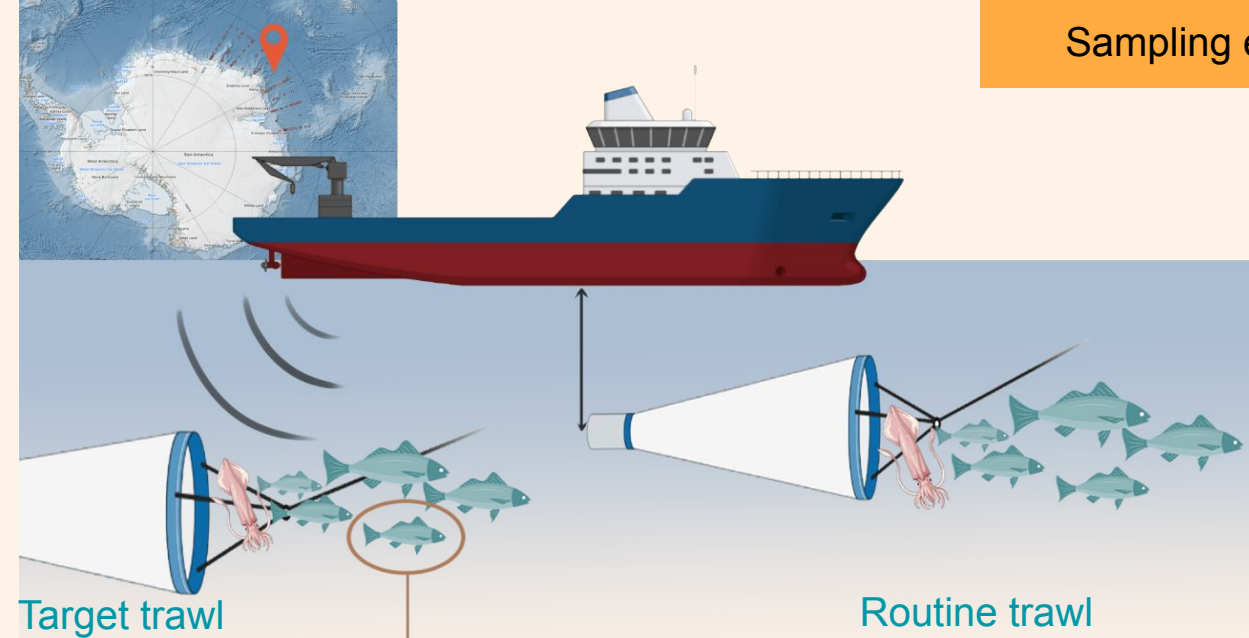
lifeStage	organismRemarks
chicks	chicks are age 0 juveniles but older than newborns

Use full name & full link in recordedBy(ID), identifiedBy(ID)

recordedBy	recordedByID	
Yi-Ming Gan	https://orcid.org/0000-0001-7087-2646	Recommended
YM Gan	0000-0001-7087-2646	common name without recordedByID is confusing, ORCID is implicit here

Sampling event core

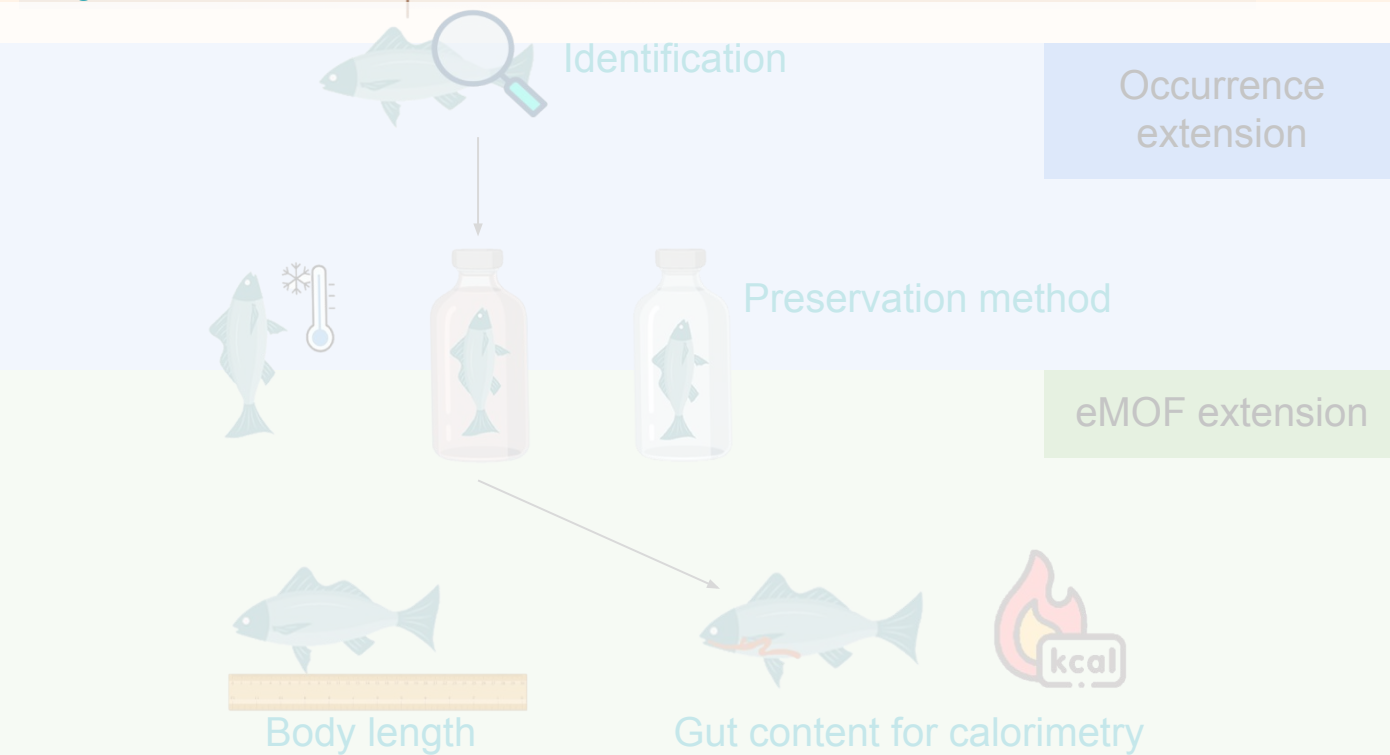
An **action** that occurs at some **location** during some **time**.



Station grid 2006-01-17/2006-03-03
 POLYGON(30 -69, 30 -32, 147 -32, 147 -69)

Sampling station with target trawl

Sampling station with routine trawl

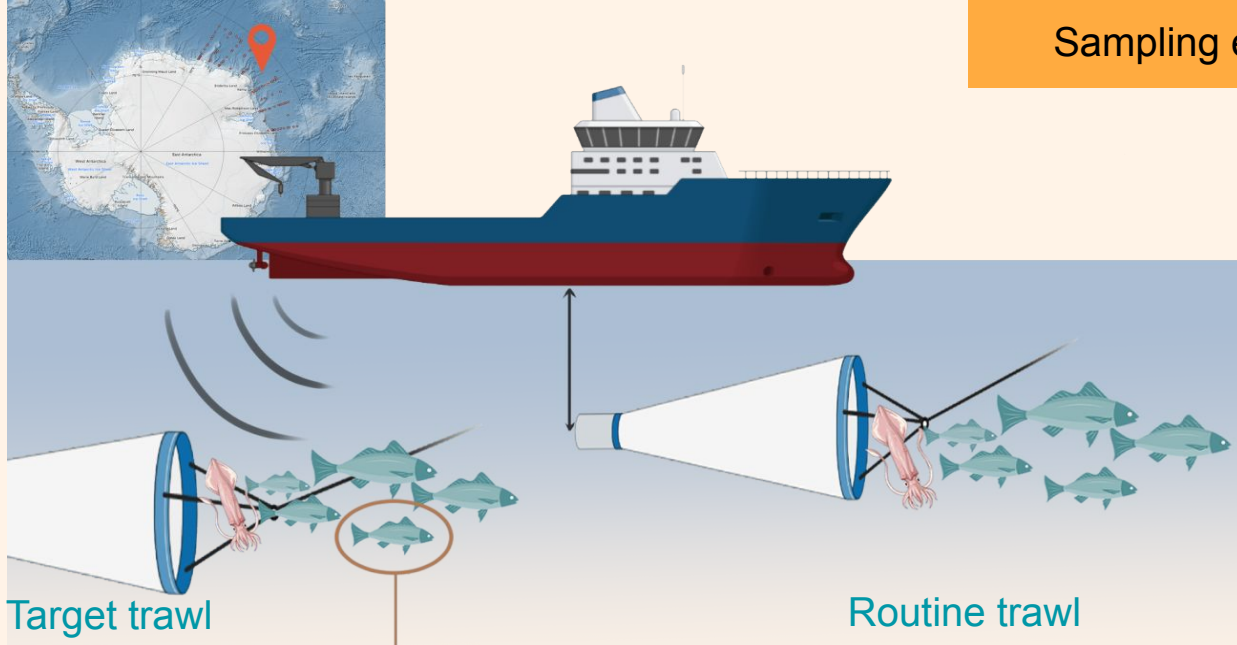


scientificName	occurrence Status	preparations
Notolepis coatsi	Present	Formalin

measurementType	measurement Value	measurement Unit
Sample volume	30320	m ³
Body length	67	mm
Energy content dry weight	25.88	KJ/g

Sampling event core

An **action** that occurs at some **location** during some **time**.



Station grid 2006-01-17/2006-03-03
 POLYGON(30 -69, 30 -32, 147 -32, 147 -69)

Sampling station with target trawl

Sampling station with routine trawl



Identification

Occurrence extension

An existence of an **Organism** at a particular **place** at a particular **time** (sampling event)



Preservation method

eMOF extension

scientificName	occurrence Status	preparations
Notolepis coatsi	Present	Formalin



Body length

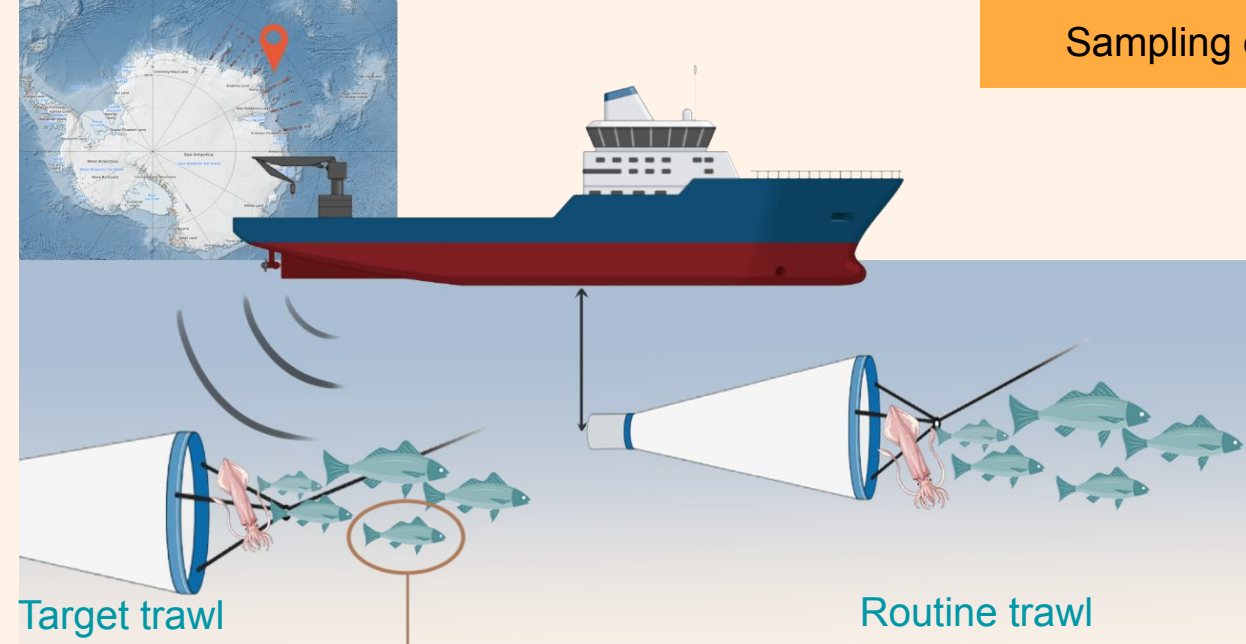


Gut content for calorimetry

measurementType	measurement Value	measurement Unit
Sample volume	30320	m ³
Body length	67	mm
Energy content dry weight	25.88	KJ/g

Sampling event core

An **action** that occurs at some **location** during some **time**.



Station grid 2006-01-17/2006-03-03
 POLYGON(30 -69, 30 -32, 147 -32, 147 -69)

Sampling station with target trawl

Sampling station with routine trawl



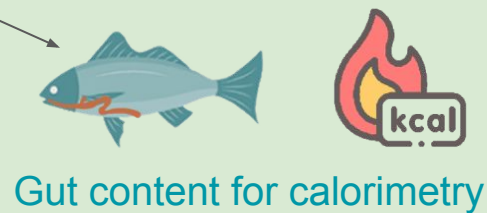
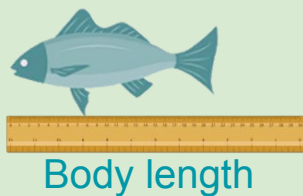
Occurrence extension

An existence of an **Organism** at a particular **place** at a particular **time** (sampling event)



eMOF extension

scientificName	occurrence Status	preparations
Notolepis coatsi	Present	Formalin



measurementType	measurement Value	measurement Unit
Sample volume	30320	m ³
Body length	67	mm
Energy content dry weight	25.88	KJ/g