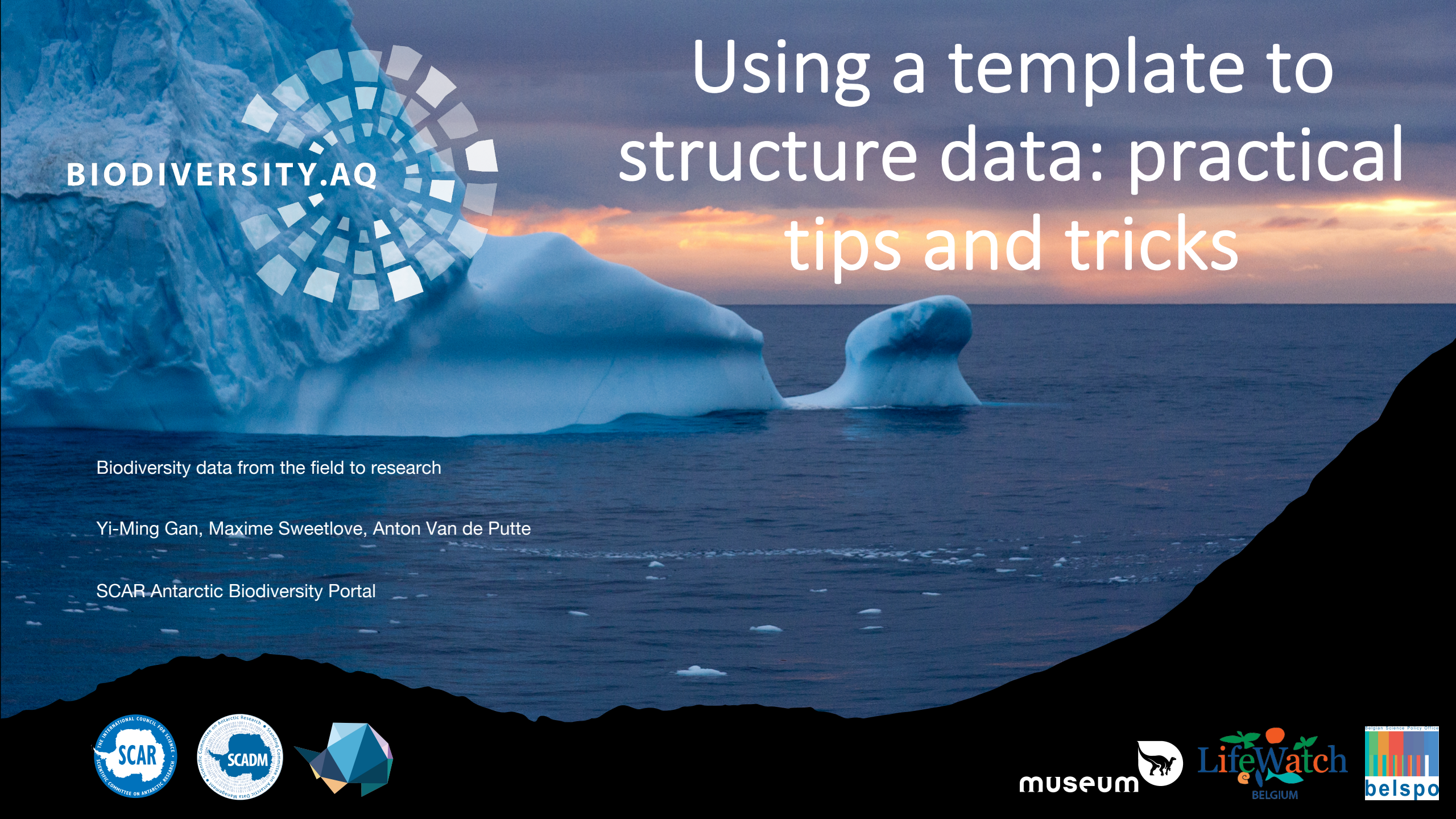


Using a template to structure data: practical tips and tricks

Biodiversity data from the field to research

Yi-Ming Gan, Maxime Sweetlove, Anton Van de Putte

SCAR Antarctic Biodiversity Portal

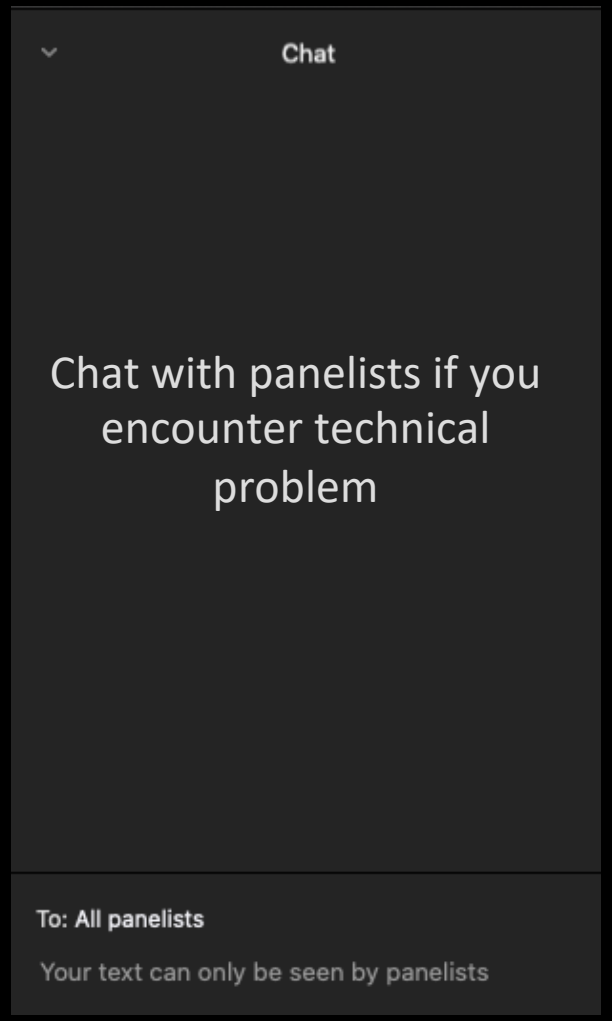
We are recording this seminar

Let us know if you prefer not to be recorded.

Code of Conduct

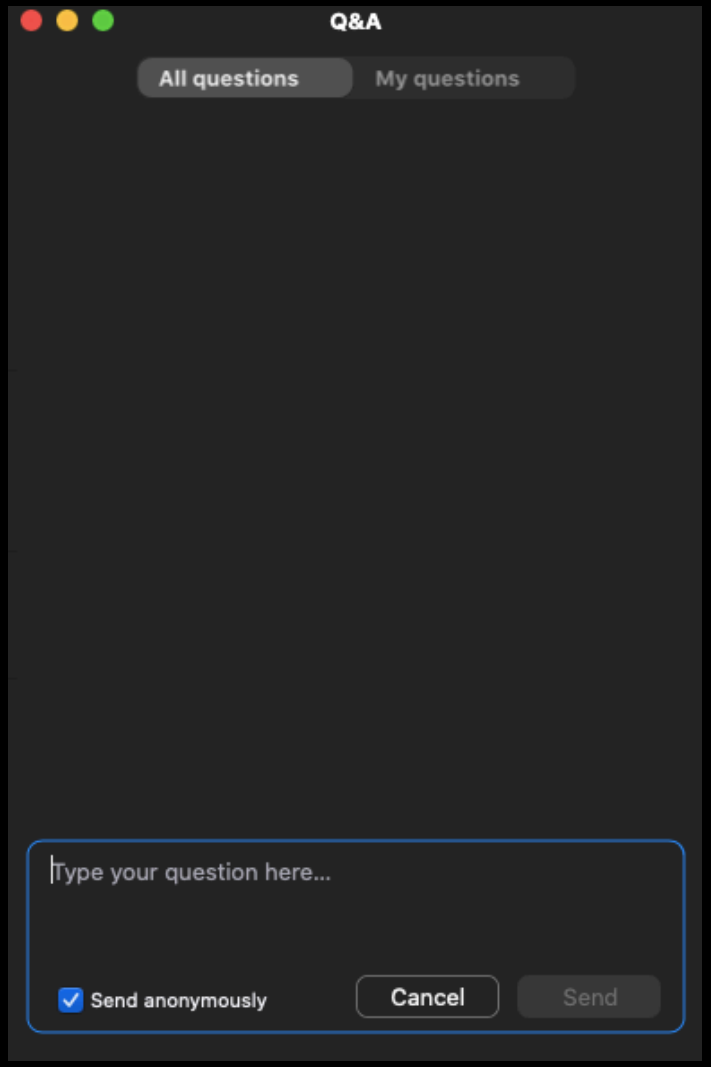
- Be respectful
- We will follow the principles of the rOpenSci Code of Conduct
  - https://ropensci.org/code-of-conduct/

Using Zoom in this webinar



Chat with panelists if you encounter technical problem

Ask questions using Q&A feature
or
raise your hand

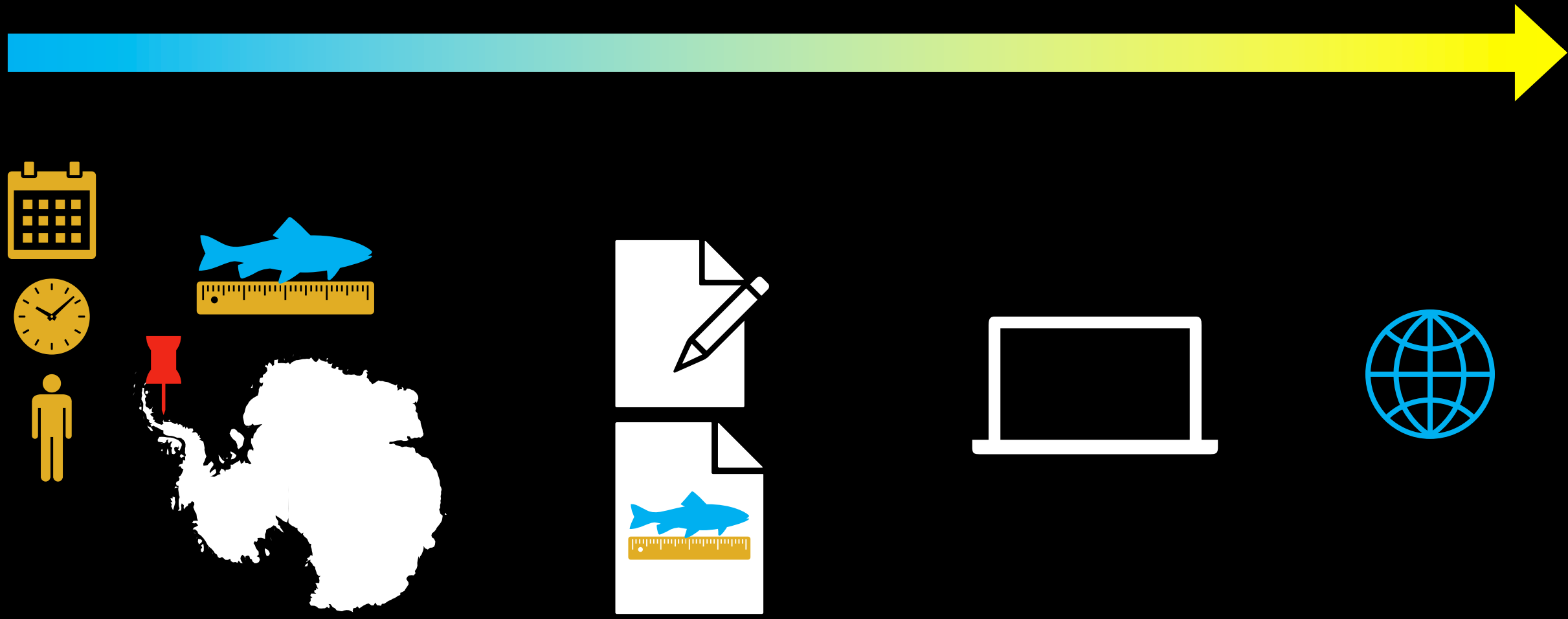


These webinars won’t cover everything

- Discover more based on the links that we provide
- Schedule a session to discuss directly
  - https://doodle.com/meetme/qc/XxYYpJwmbG

- Physical observation to digital data
- Example: sampling at sea
- Example: genomics data

Physical observation to digital data



Physical observation

Physical records

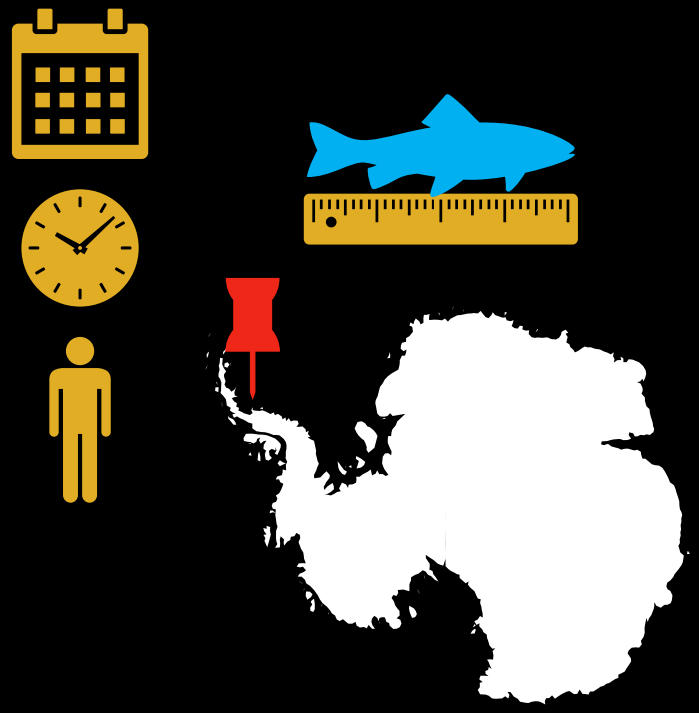
Digital record

Distributed data

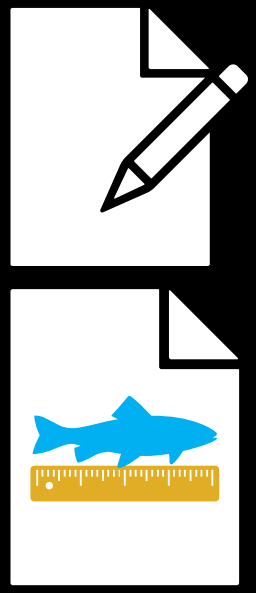
Physical observation to digital data

Human friendly

Machine friendly



Physical observation

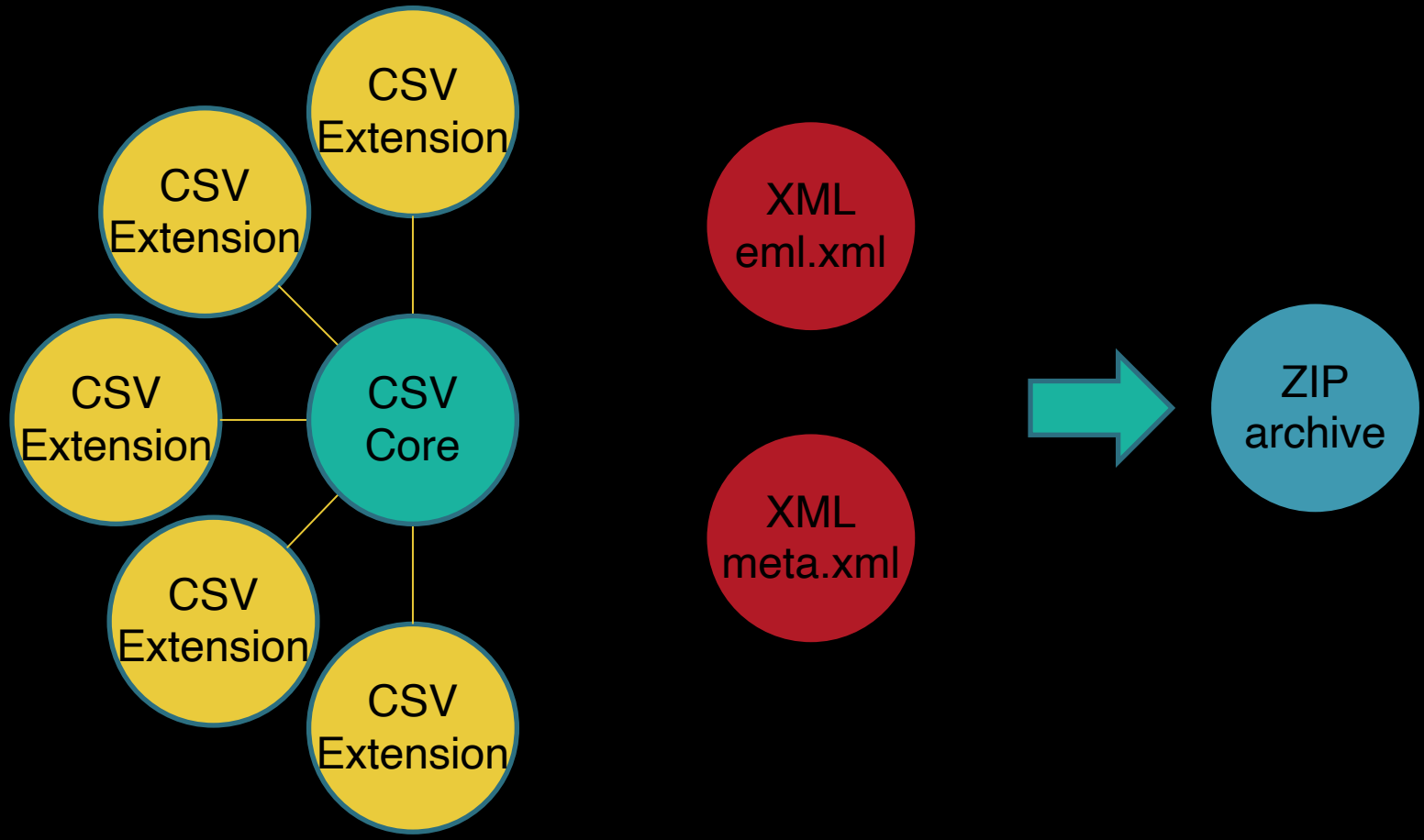


Physical records

Digital record

Distributed data

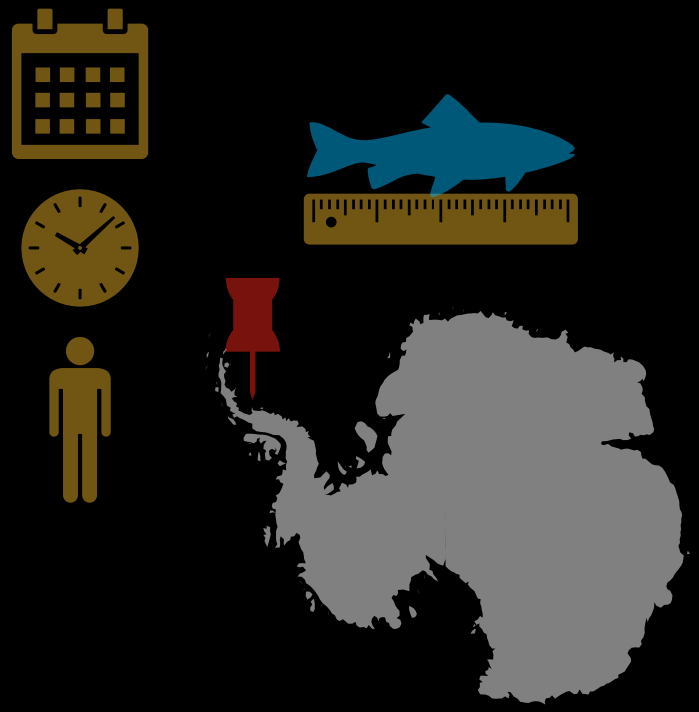
Darwin Core standard serves to distribute biodiversity data



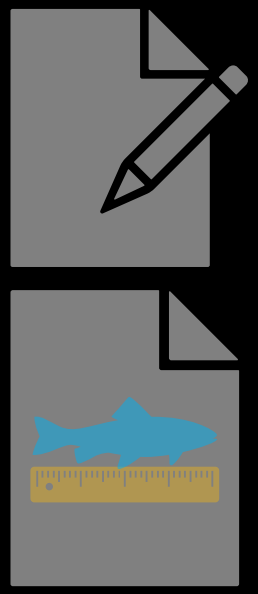
Darwin Core standard serves to distribute biodiversity data

Human friendly

Machine friendly



Physical observation



Physical records

Digital record

Darwin Core

Distributed data

Darwin Core quick reference guide

https://dwc.tdwg.org/terms/

Darwin Core extensions

https://tools.gbif.org/dwca-validator/extensions.do

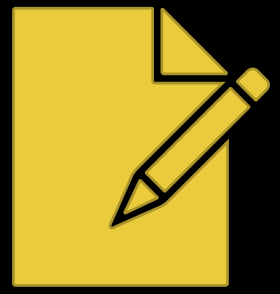
Darwin core validator

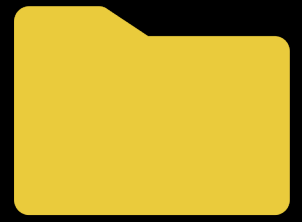
https://www.gbif.org/tools/data-validator/about

Darwin core instructions

https://github.com/gbif/ipt/wiki/howToPublish#instructions
https://obis.org/manual/ipt/

Objectives





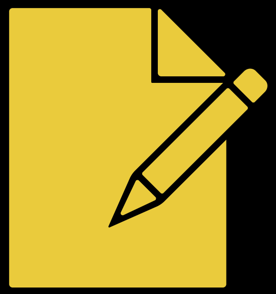
field templates

- Design to help scientists record and aggregate data
- Customized according to needs (1-1 feedback sessions)
- Human readable and writeable

Data templates

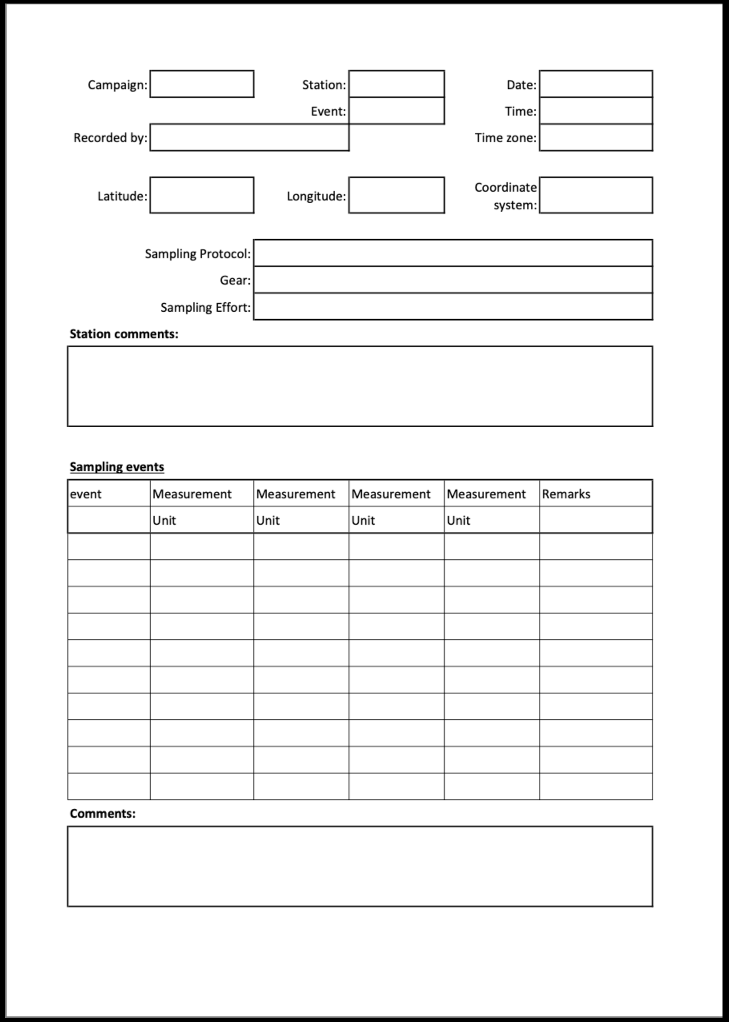
- Design for data exchange across the web
- Allows integrating data from different field expeditions
- Data standard
- Machine readable

2 types of templates



Field templates

- Record data in the **field**
- Per sampling station/samples



Campaign:
Station:
Event:
Date:
Time:
Recorded by:
Time zone:
Latitude:
Longitude:
Coordinate system:
Sampling Protocol:
Gear:
Sampling Effort:

Station comments:

Sampling events

event	Measurement	Measurement	Measurement	Measurement	Remarks
	Unit	Unit	Unit	Unit	

Comments:

Campaign:
Recorded by:
Station number:
Event ID:
Date:

Summary

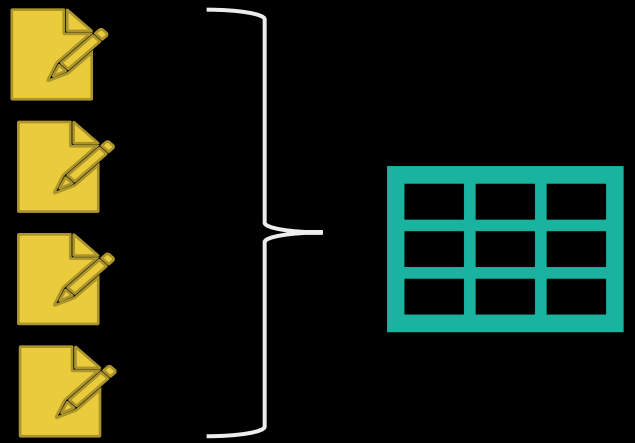
Sample ID	Scientific Name	Count	Measurement	Measuremer	Measuremer	Boolean field	Remarks
		Individual	Unit	Unit	Unit	Y/N	

Remarks

2 types of templates

Field	occurrenceID	basisOfRecord	decimalLongitude	decimalLatitude	*geodeticDatum*	eventDate	occurrenceStatus	scientificName	scientificNameID
Definition, see https://dwc.tdwg.org/terms/	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.	The specific nature of the data record.	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.	The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context.	A statement about the presence or absence of a Taxon at a Location.	The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.	An identifier for the nomenclatural (not taxonomic) details of a scientific name.
Remarks	Mandatory field for GBIF and OBIS, unique per record	Mandatory field for GBIF and OBIS, standard label only. This column contains validation of vocabularies for basisOfRecord stated in "vocabulary" sheet.	Mandatory field for OBIS. This column contains data validation. Only decimal values of -180 and 180 are valid	Mandatory field for OBIS. This column contains data validation. Only decimal values of -90 and 90 are valid	Recommended best practice is to use the EPSG code of the SRS	Mandatory field for OBIS. Recommended best practice is to use a date that conforms to ISO 8601-1:2019.	Mandatory field for OBIS.	Mandatory field for OBIS.	Mandatory field for OBIS.
Example	urn:catalog:UWBM:Bird:89776	PreservedSpecimen	-121.17611	-41.09834	EPSG:4326	2020-10-31	presence, absence	Ctenomys sociabilis (genus + specificEpithet)	urn:lsid:ipni.org:names:37829-1:1.3

Field	**eventID**	parentEventID	**eventDate**	samplingProtocol	fieldNumber	**decimalLatitude**	**decimalLongitude**	geodeticDatum	footprintWKT	eventRemarks
Definition (see more https://dwc.tdwg.org/terms/)	An identifier for the set of information associated with an Event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the data set.	An identifier for the broader Event that groups this and potentially other Events.	The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context.	The name of, reference to, or description of the method or protocol used during an Event.	An identifier given to the event in the field. Often serves as a link between field notes and the Event.	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.	A Well-Known Text (WKT) representation of the shape (footprint, geometry) that defines the Location. A Location may have both a point-radius representation (see decimalLatitude) and a footprint representation, and they may differ from each other.	Comments or notes about the Event.
Comments			Mandatory field for OBIS. Recommended best practice is to use a date that conforms to ISO 8601-1:2019.			Mandatory field for OBIS.	Mandatory field for OBIS.			
Examples	INBO:VIS:Ev:00009375	A1 (parentEventID to identify the main Whittaker Plot in nested samples, each with its own eventID - A1:1, A1:2).	2020-10-31			-41.09834	-121.17611	EPSG:4326	POLYGON ((10 20, 11 20, 11 21, 10 21, 10 20)) (the one-degree bounding box with opposite corners at longitude=10, latitude=20 and longitude=11, latitude=21)	After the recent rains the river is nearly at flood stage.



Data templates

- Aggregate data from multiple field templates
- Multiple stations/samples

On identifiers

- Don’t change them
- If you can keep them logical
- Don’t use strange characters
- Did I say don’t change them?
- An example!!!

Example: sampling at sea

Original identifiers

RV105_001

RV105_001_01

RV105_001_02

RV105_001_03

RV105_002

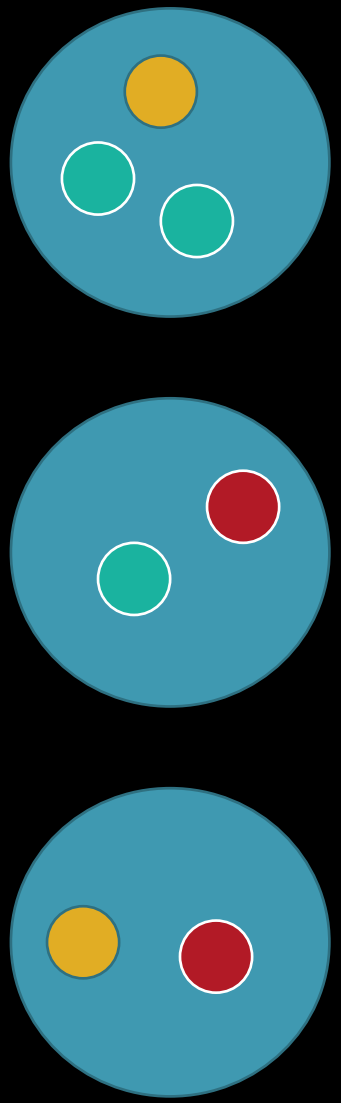
RV105_002_01

RV105_002_02

RV105_003

RV105_003_01

RV105_003_02



Identifiers people come up with

CTD_01

Trawl_01

Trawl_02

Sample_01_01

Sample_02_01

Run1

01_1

Run2

01_2

Trawl_03

Trawl_01

Sample_1

Sample_2

Trawl_02

Sample_3

CTD_02

Original identifiers

RV105_001

RV105_001_01

RV105_001_02

RV105_001_03

RV105_002

RV105_002_01

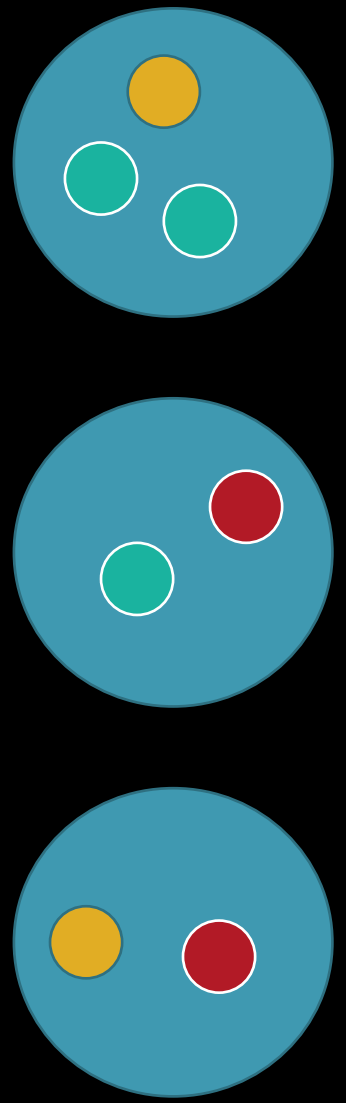
RV105_002_02

RV105_003

RV105_003_01

RV105_003_02

Identifiers people come up with



CTD_01

Trawl_01

Trawl_02

Trawl_03

Trawl_01

Trawl_02

CTD_02

Sample_01_01

Sample_02_01

Sample_1

Sample_2

Sample_3

Run1

01_1

Run2

01_2

Field Event template

Campaign:		Station:		Date:	
		Event:		Time:	
Recorded by:				Time zone:	
Latitude:		Longitude:		Coordinate system:	

Sampling Protocol:	
Gear:	
Sampling Effort:	

Station comments:

Sampling events

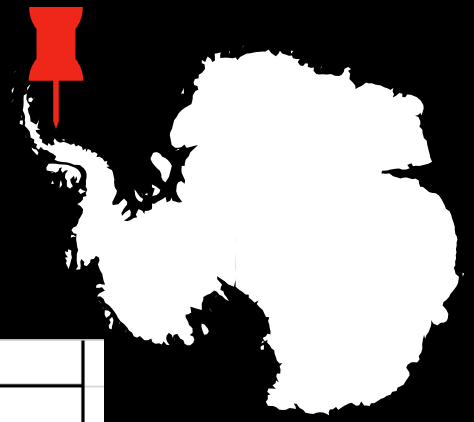
event	Measurement	Measurement	Measurement	Measurement	Remarks
	Unit	Unit	Unit	Unit	

Comments:

How are the templates organized?

The field **event** sheet for sampling events data

Sampling station



Sampling event 1

Sampling event 2

Sub-event 1

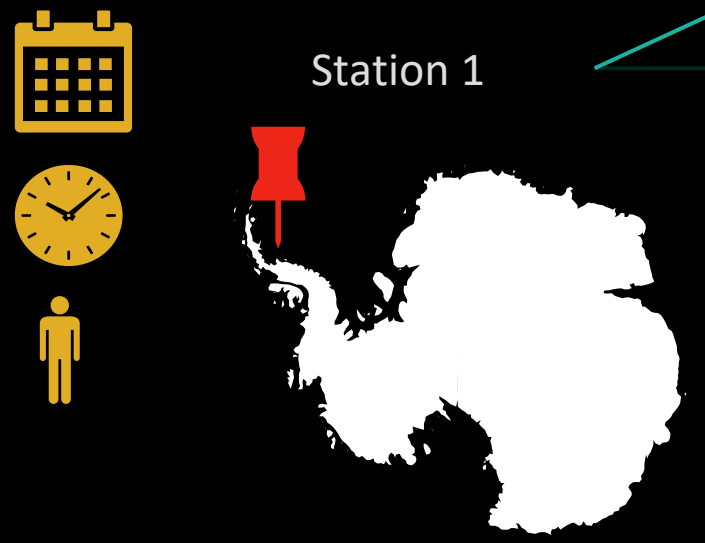
Sub-event 2

Campaign:		Station:		Date:	
		Event:		Time:	
Recorded by:				Time zone:	
Latitude:		Longitude:		Coordinate system:	
Sampling Protocol:					

Sampling events					
event	Measurement	Measureme	Measurement	Measuremen	Remarks
	Unit	Unit	Unit	Unit	

How are the templates organized?

The field **event** sheet for sampling events data



Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Campaign:		Station number:		Event ID:	
Recorded by:				Date:	

Summary

Sample ID	Scientific Name	Count	Measurement	Measuremer	Measuremer	Boolean field	Remarks
		Individual	Unit	Unit	Unit	Y/N	

Remarks

Field occurrence template

How are the templates organized?

The field **occurrence** sheet for sampling events data

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net

RMT Net 1

RMT Net 2

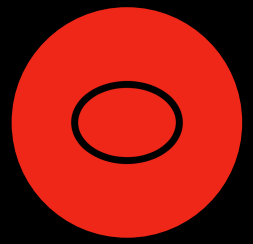
Jar 01

Jar 02

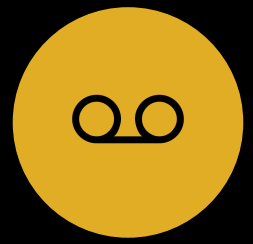
Field occurrence sheet
for **RMT Net 1**

Sample ID	Species	Count
Jar 01		1
Jar 02		2

WHAT

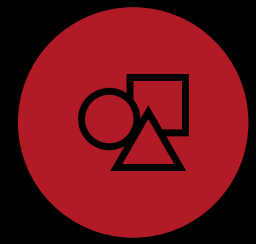


WHERE



WHEN

WHO



HOW

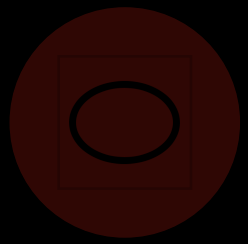
Field	occurrenceID	basisOfRecord	decimalLongitude	decimalLatitude	*geodeticDatum*	eventDate	occurrenceStatus	scientificName	scientificNameID
Definition, see https://dwc.tdwg.org/terms/	An identifier for the Occurrence (as opposed to a particular digital record of the occurrence). In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the occurrenceID globally unique.	The specific nature of the data record.	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.	The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context.	A statement about the presence or absence of a Taxon at a Location.	The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic rank that can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.	An identifier for the nomenclatural (not taxonomic) details of a scientific name.
Remarks	Mandatory field for GBIF and OBIS, unique per record	Mandatory field for GBIF and OBIS, standard label only. This column contains validation of vocabularies for basisOfRecord stated in "vocabulary" sheet.	Mandatory field for OBIS. This column contains data validation. Only decimal values of -180 and 180 are valid	Mandatory field for OBIS. This column contains data validation. Only decimal values of -90 and 90 are valid	Recommended best practice is to use the EPSG code of the SRS	Mandatory field for OBIS. Recommended best practice is to use a date that conforms to ISO 8601-1:2019.	Mandatory field for OBIS.	Mandatory field for OBIS.	Mandatory field for OBIS.
Example	urn:catalog:UWBM:Bird:89776	PreservedSpecimen	121.17611	41.09834	EPSG:4326	2020-10-31	presence, absence	Ctenomys sociabilis (genus + specificEpithet)	urn:lsid:ipni.org:names:37829-1:1.3

Data occurrence template

Occurrence Core: Key terms

occurrenceID	**basisOfRecord**	**scientificName**	**scientificNameID**
occurrenceStatus	**decimalLongitude**	**decimalLatitude**	**eventDate**

WHAT



WHERE

WHEN

WHO

HOW

occurrenceID

occurenceID: example

occurenceID	institutionCode	collectionCode	catalogNumber
http://arctos.database.museum/guid/MSB:Mamm:233627	MSB	Mamm	233627
PS89_FF_000023	PS89	FF	000023

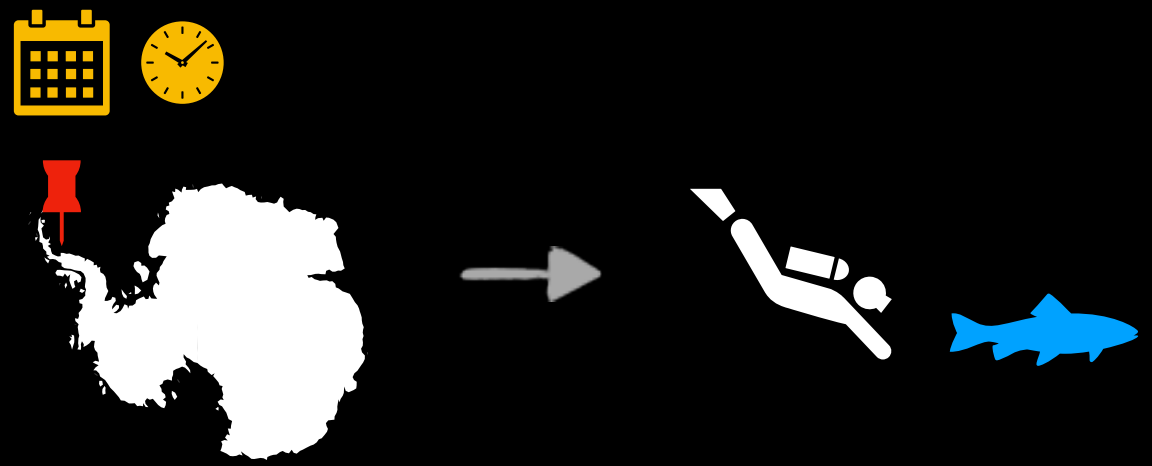
basisOfRecord

Mandatory term with standard labels

- HumanObservation
- MachineObservation
- PreservedSpecimen
- MaterialSample
- LivingSpecimen
- FossilSpecimen

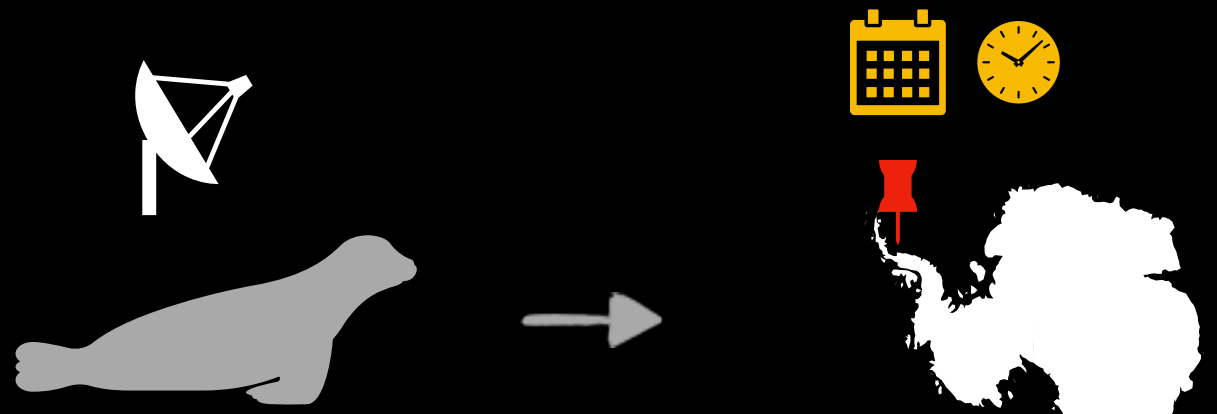
Basis of record

HumanObservation



Basis of record

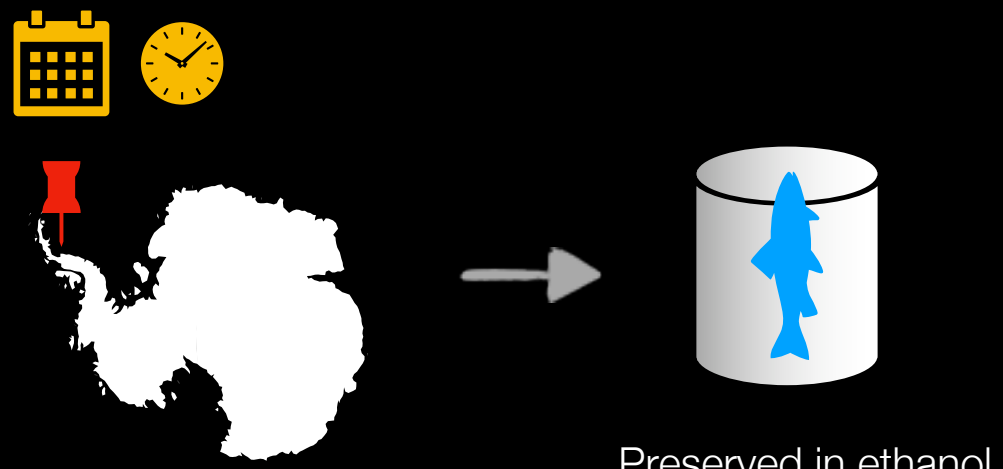
MachineObservation



GPS tracker

Basis of record

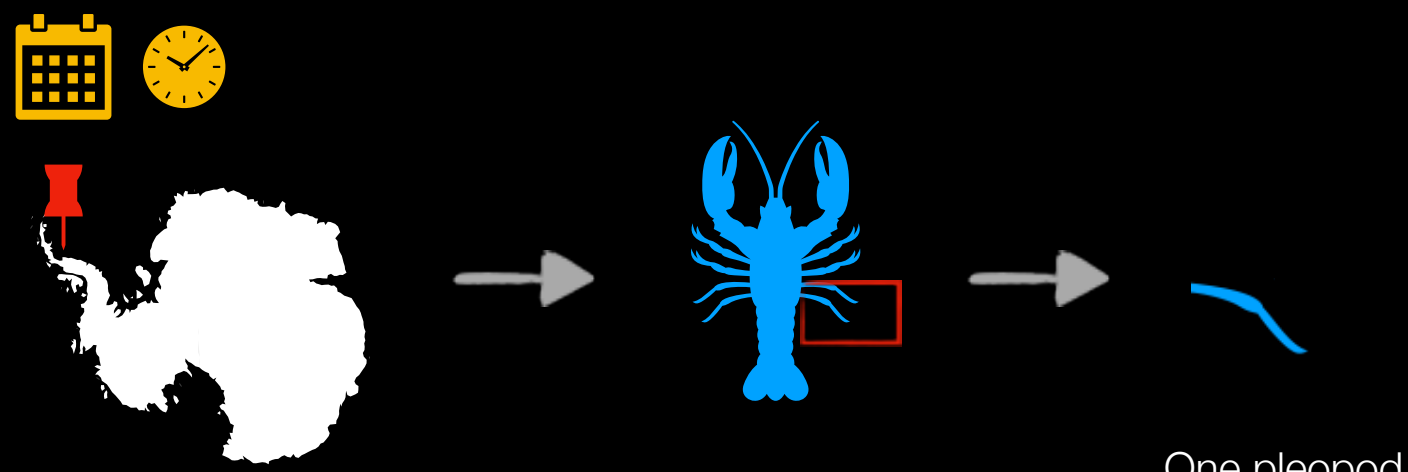
PreservedSpecimen



Preserved in ethanol

Basis of record

MaterialSample



One pleopod

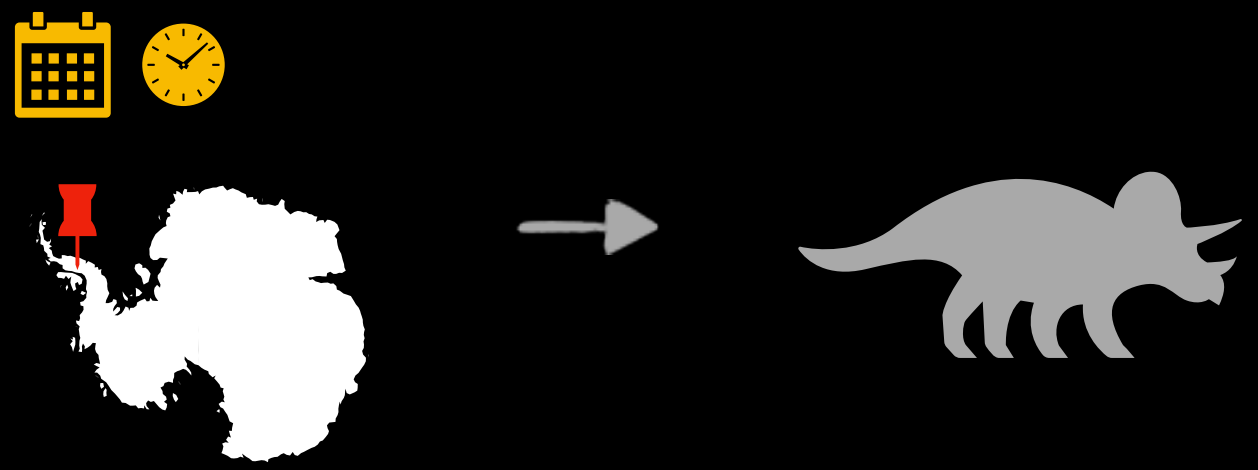
Basis of record

LivingSpecimen

A living plant in a botanical garden

Basis of record

FossilSpecimen



BasisOfRecord: Provide context in combination of other fields

basisOfRecord	• HumanObservation
type	• StillImage
associatedMedia	• Link to media (doi)

This record is an observation made by human based on still image from the associated media

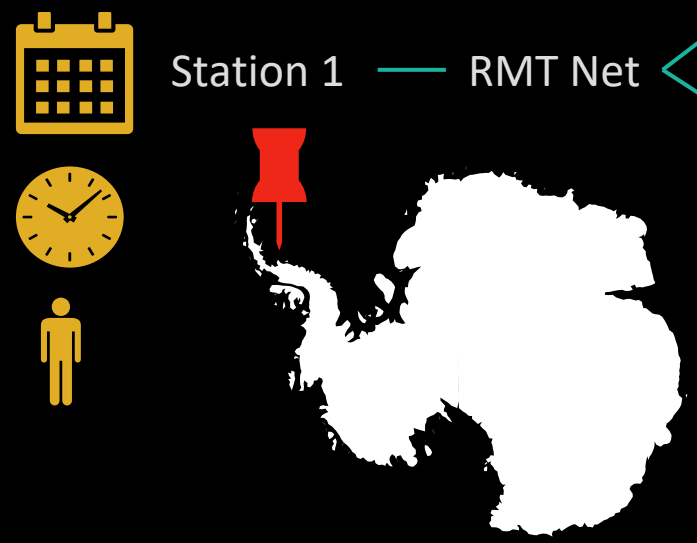
IndivudualCount

- Not A required term for OBIS/GBIF
- Highly recommended by us

Filling the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200



Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	Species	individualCount	basisOfRecord
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen

scientificName

scientificNameID

WoRMS	Global Names Index (GNI)	Integrated Taxonomic Information System (ITIS)
• Register of Antarctic (marine) Species		

scientificName, scientificNameID

Taxonomy

Uncertain identifications

Reduce taxonomy to first common higher classification level

cf.
- *confer*
- resembles, looks like

aff.
- *affinis*
- related to but not identical to

Species *Electrona antarctica*

Species *Electrona aff. antarctica*

Vandepitte, L. (2017) Biodiversity data standards: Taxonomy. OBIS nodes training course. https://classroom.oceanteacher.org/mod/resource/view.php?id=7439

Taxonomy

Uncertain identifications

Reduce taxonomy to first common higher classification level

Species *Acanthonchus duplicatus* <> Species *Acanthonchus cf. duplicatus*

scientificName	genus	identificationQualifier	taxonRank
Acanthonchus	Acanthonchus	cf. duplicatus	genus

Vandepitte, L. (2017) Biodiversity data standards: Taxonomy. OBIS nodes training course. https://classroom.oceanteacher.org/mod/resource/view.php?id=7439

Examples

scientificName	scientificName Authorship	Identification Qualifier	taxonrank	scientificNameID
Electrona	Goode & Bean, 1896	aff. risso	genus	urn:lsid:marinespecies.org:taxname:125821)
Electrona antarctica	Günther, 1878		species	(urn:lsid:marinespecies.org:taxname:217697)
Electrona	Goode & Bean, 1896	sp.	genus	urn:lsid:marinespecies.org:taxname:125821)
Electrona	Goode & Bean, 1896	spp.	genus	urn:lsid:marinespecies.org:taxname:125821)

identifiedBy
dateIdentified

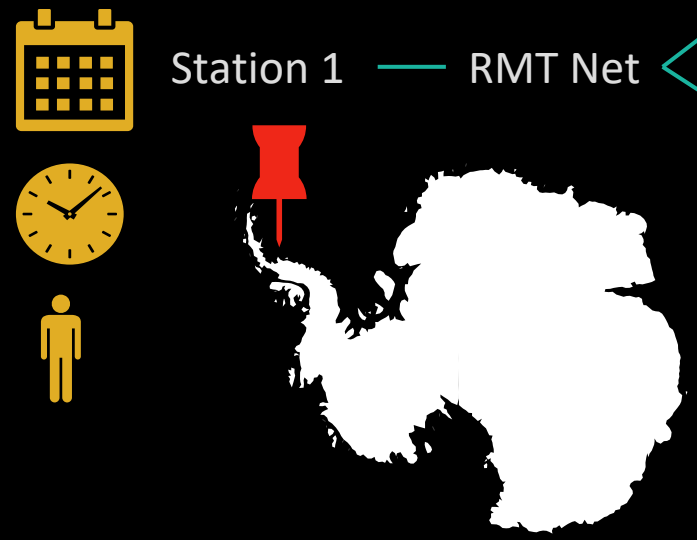
Acknowledge the personnel who carry out different tasks:
e.g. Taxonomist who identified the species
(https://bionomia.net/)

WHAT WHERE WHEN WHO HOW

Filling the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200




Jar 01


Jar 02

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	ScientificName	individualCount	basisOfRecord	scientificNameID	identifiedBy	IdentifiedDate
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	125821	Anton Van de Putte	2008-11-05
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	217697	Anton Van de Putte	2008-11-05

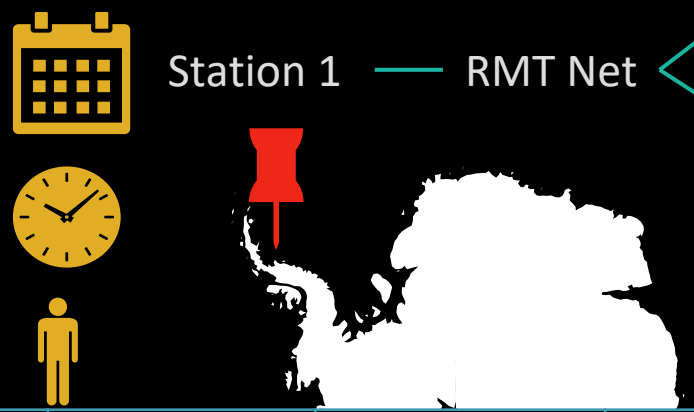
occurrenceStatus

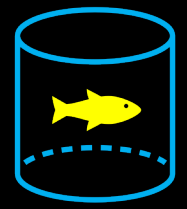
Present/absent

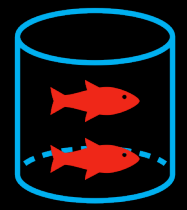
Filling the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200




Jar 01


Jar 02

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

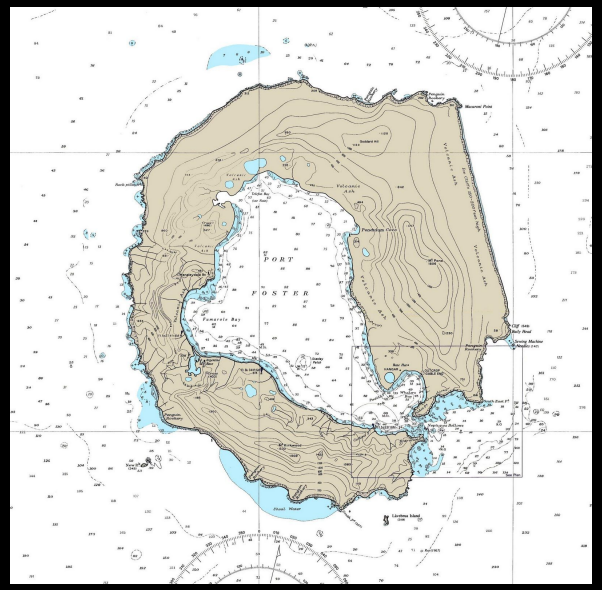
occurrenceID	ScientificName	individualCount	basisOfRecord	scientificNameID	Occurrence status
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	125821	present
ANT1_Stn1_RMT-Net_1_Jar01		0	HumanObservation	217697	absent
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	217697	present
ANT1_Stn1_RMT-Net_1_Jar02		0	HumanObservation	125821	absent

Recommended information

WHAT

- *organismQuantity*
- *organismQuantityType*
- *sex*
- *lifeStage*
- *behavior*
- *occurrenceRemarks*

WHAT WHERE WHEN WHO HOW



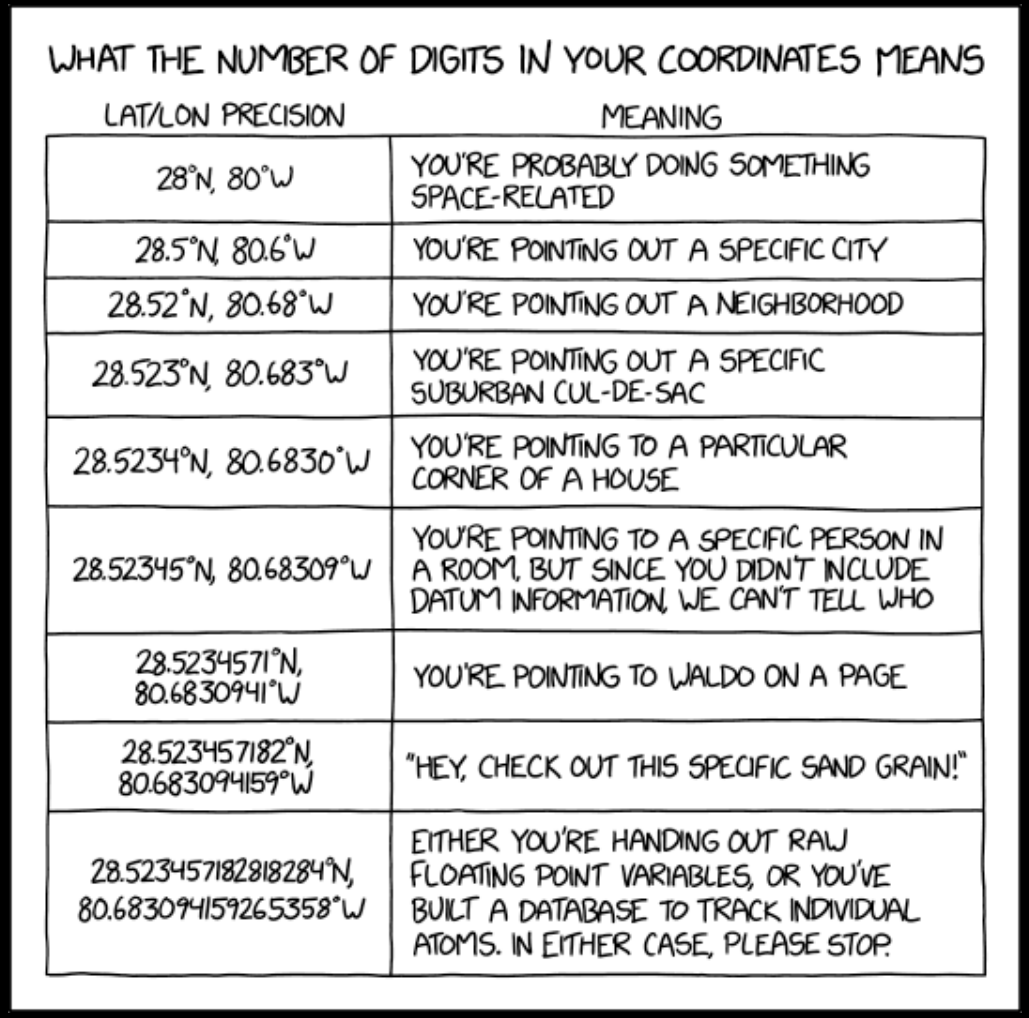
Geographic coordinates

decimalLongitude

decimalLatitude

Geographic coordinates

Coordinates precision



WHAT THE NUMBER OF DIGITS IN YOUR COORDINATES MEANS

LAT/LON PRECISION	MEANING
28°N, 80°W	YOU'RE PROBABLY DOING SOMETHING SPACE-RELATED
28.5°N, 80.6°W	YOU'RE POINTING OUT A SPECIFIC CITY
28.52°N, 80.68°W	YOU'RE POINTING OUT A NEIGHBORHOOD
28.523°N, 80.683°W	YOU'RE POINTING OUT A SPECIFIC SUBURBAN CUL-DE-SAC
28.5234°N, 80.6830°W	YOU'RE POINTING TO A PARTICULAR CORNER OF A HOUSE
28.52345°N, 80.68309°W	YOU'RE POINTING TO A SPECIFIC PERSON IN A ROOM. BUT SINCE YOU DIDN'T INCLUDE DATUM INFORMATION, WE CAN'T TELL WHO
28.5234571°N, 80.6830941°W	YOU'RE POINTING TO WALDO ON A PAGE
28.523457182°N, 80.683094159°W	"HEY, CHECK OUT THIS SPECIFIC SAND GRAIN!"
28.5234571828184°N, 80.6830941592653 58°W	EITHER YOU'RE HANDING OUT RAW FLOATING POINT VARIABLES, OR YOU'VE BUILT A DATABASE TO TRACK INDIVIDUAL ATOMS. IN EITHER CASE, PLEASE STOP.

https://xkcd.com/2170/

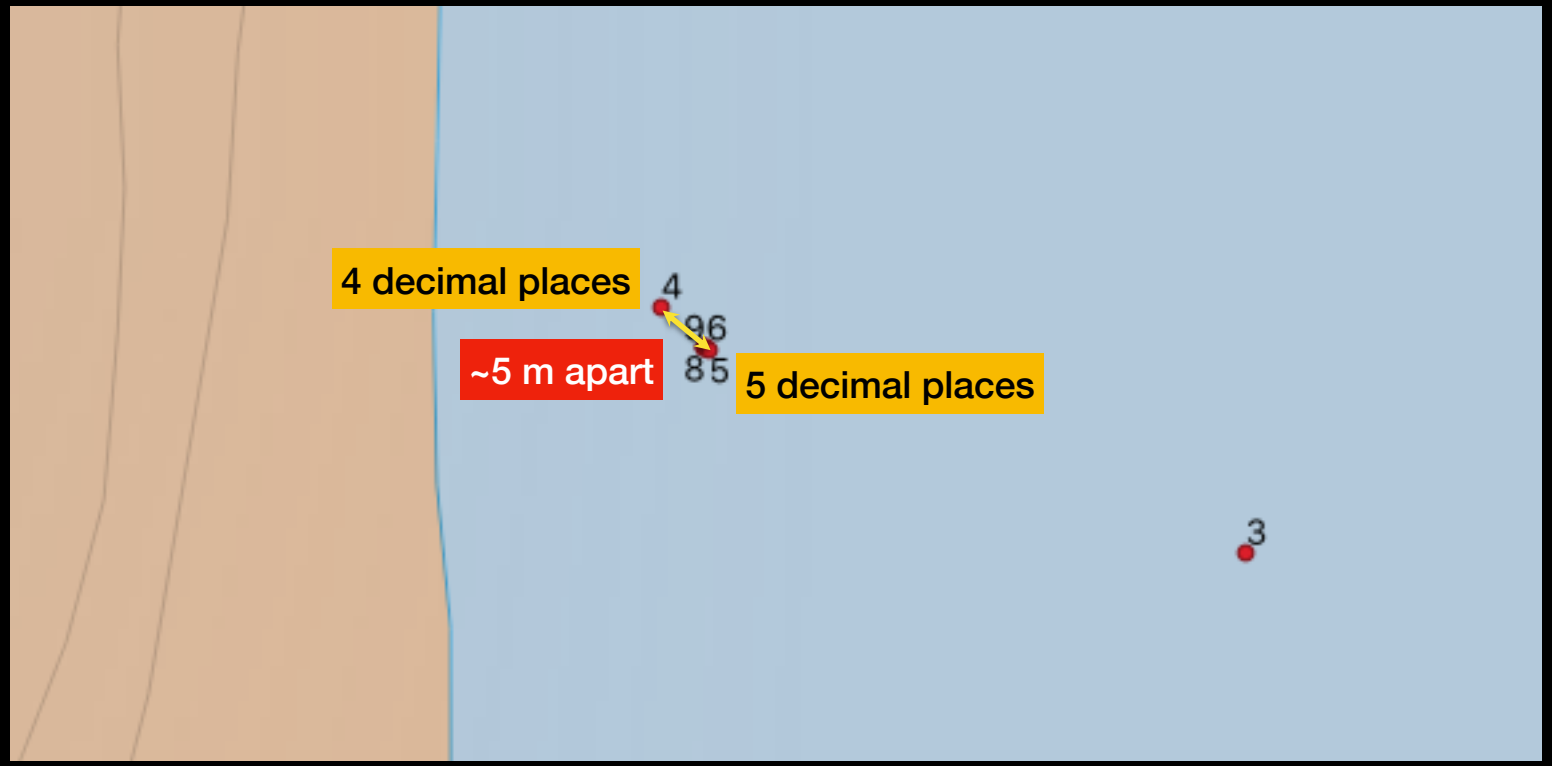
Geographic coordinates

Coordinates precision

Coordinates with many (>5) decimal places maybe resulted from data conversion
(e.g. converting degrees, minutes, seconds to decimal degrees)

Geographic coordinates

Coordinates precision



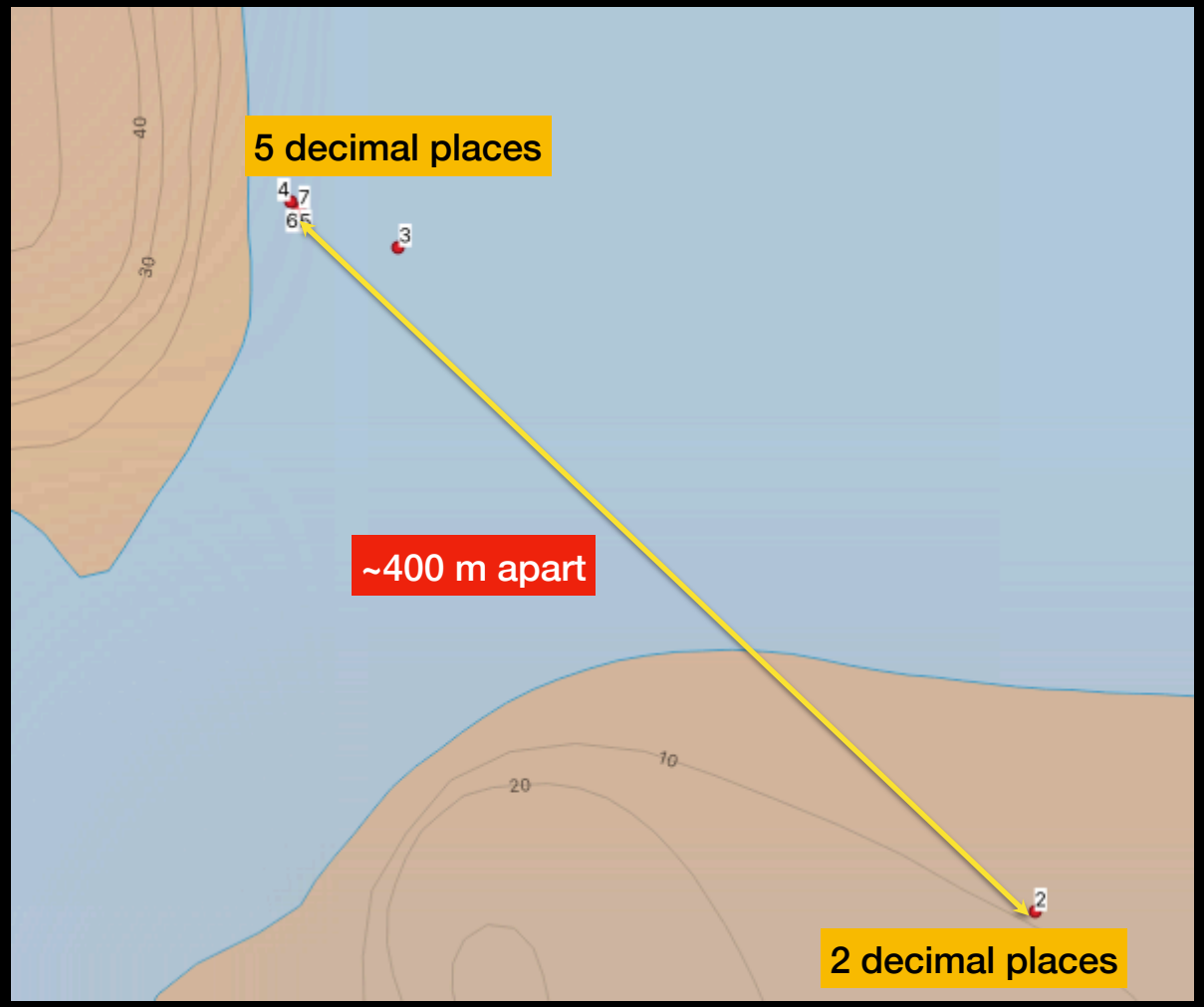
Software: QGIS 3.14, Layer: Quantarctica 3 - Detailed basemap, Antarctic polar stereographic projection (EPSG:3031), coordinates recorded in decimal degrees (EPSG:4326)

Geographic coordinates

Coordinates precision

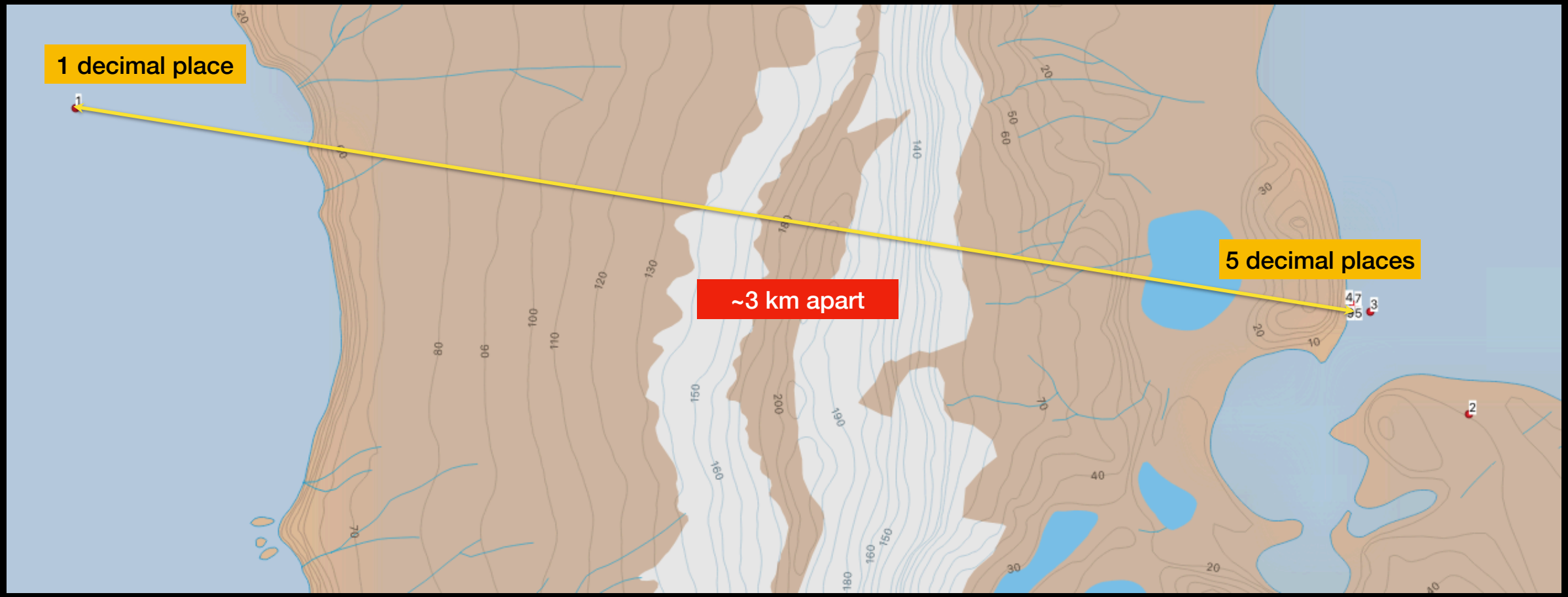
Geographic coordinates

Coordinates precision



Geographic coordinates

Coordinates precision



Geographic coordinates

Coordinates precision

What about sensitive data?

Geographic coordinates

Coordinates precision

Which precision is appropriate?

Geographic coordinates

Coordinates precision

How sensitive is your species?

To what extend your data needs to be generalized

Geographic coordinates

Coordinates precision

Chapman AD (2020) Current Best Practices for Generalizing Sensitive Species Occurrence Data (Community review draft). Version 2. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-5jp4-5g10.

Geographic coordinates

Longitude, latitude switched

Geographic coordinates

Longitude, latitude switched

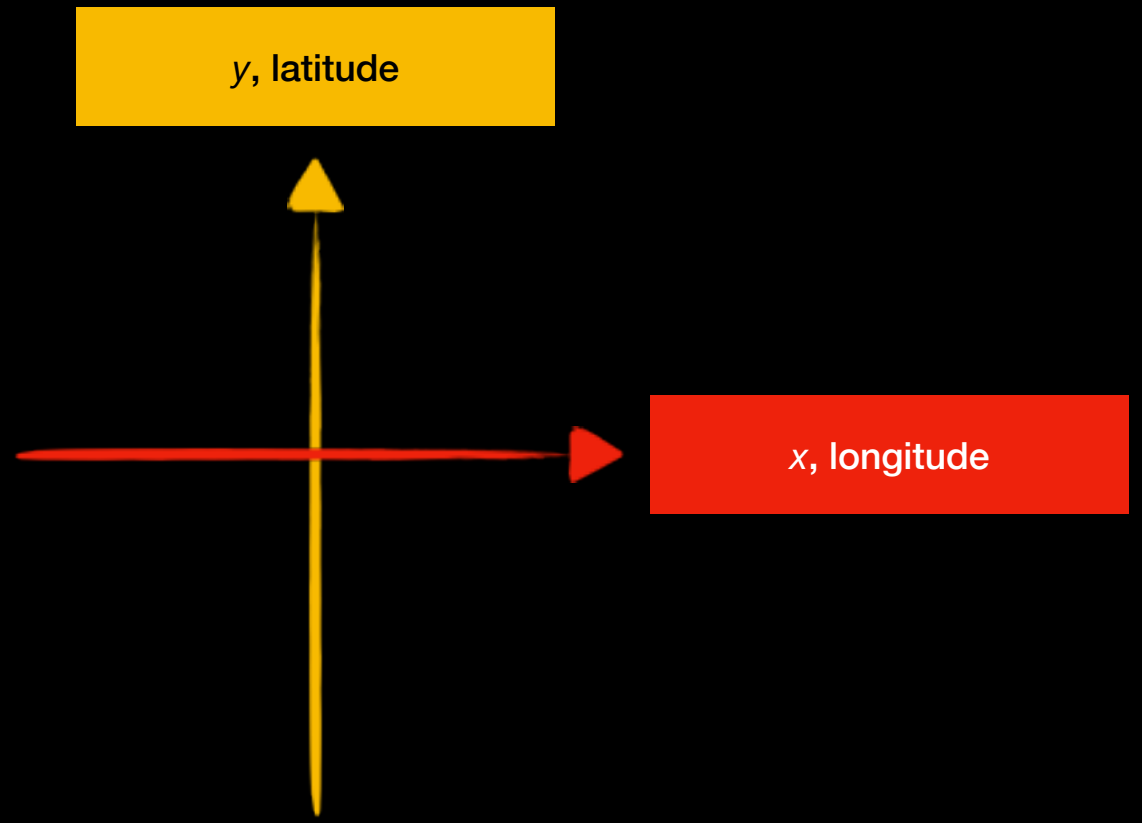
Often due to human errors when recording coordinates

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

Geographic coordinates

Longitude, latitude switched

Sometimes due to confusion

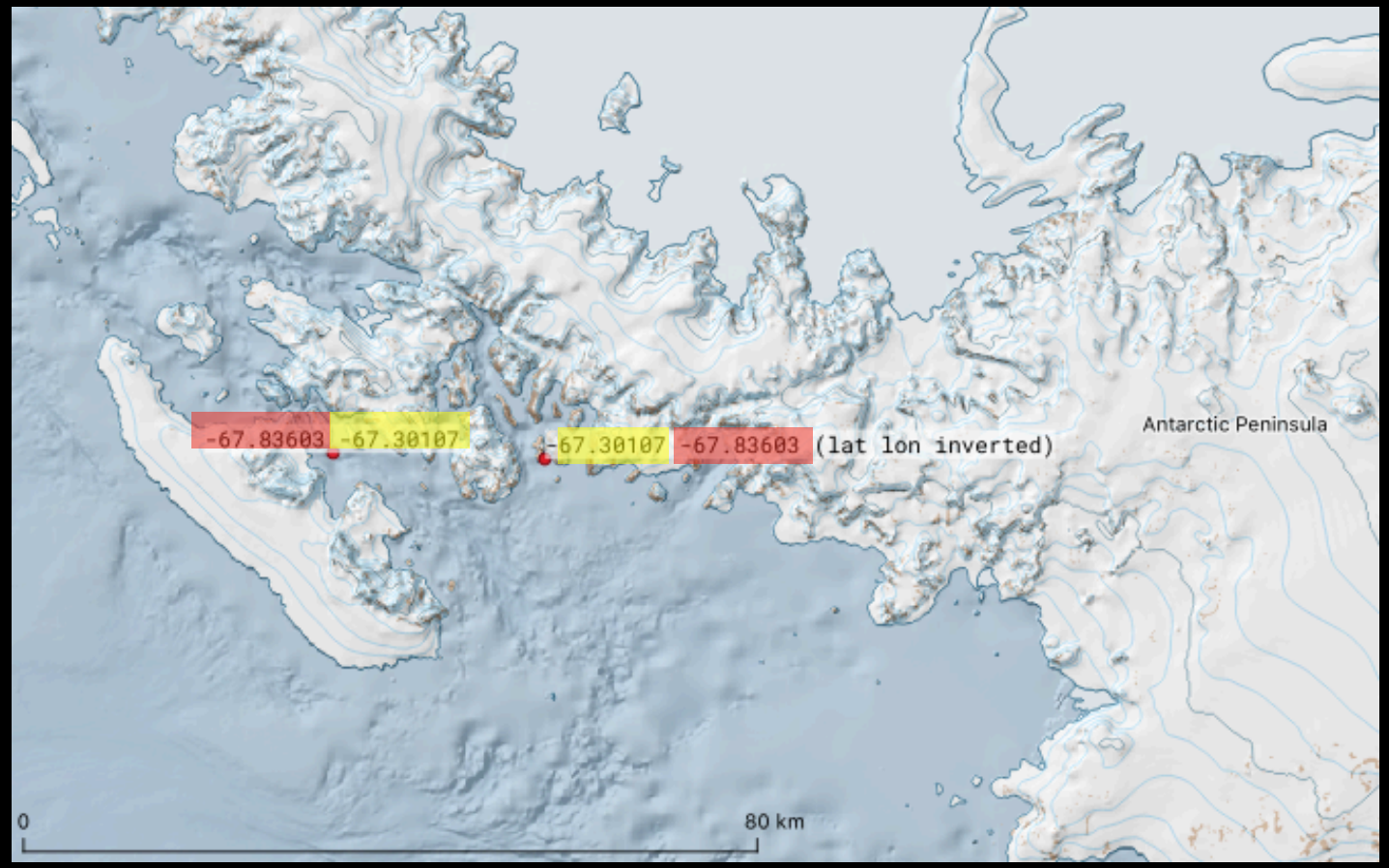


Geographic coordinates

Consequence of switched longitude and latitude

Longitude

Latitude



Marine species - Both points are in water, which is correct?

Atomize columns

Best to record latitude and longitude in 2 separate columns

verbatimLongitude	verbatimLatitude
-67° 50' 9.71"	-67° 18' 3.85"

Atomize columns

To ease downstream batch operations

verbatimLongitude	verbatimLatitude
-67° 50' 9.71"	-67° 18' 3.85"

Function can be applied to whole column

decimalLongitude	**decimal**Latitude
-67.83603	-67.30107

Atomize columns

Function can be applied to whole column

verbatimLongitude	verbatimLatitude
-67° 50' 9.71"	-67° 18' 3.85"

decimalLongitude	**decimal**Latitude
-67.83603	-67.30107

footprintWKT
POINT (-67.83603 -67.30107)

Geographic coordinates

Consistent format

Geographic coordinates

Be consistent

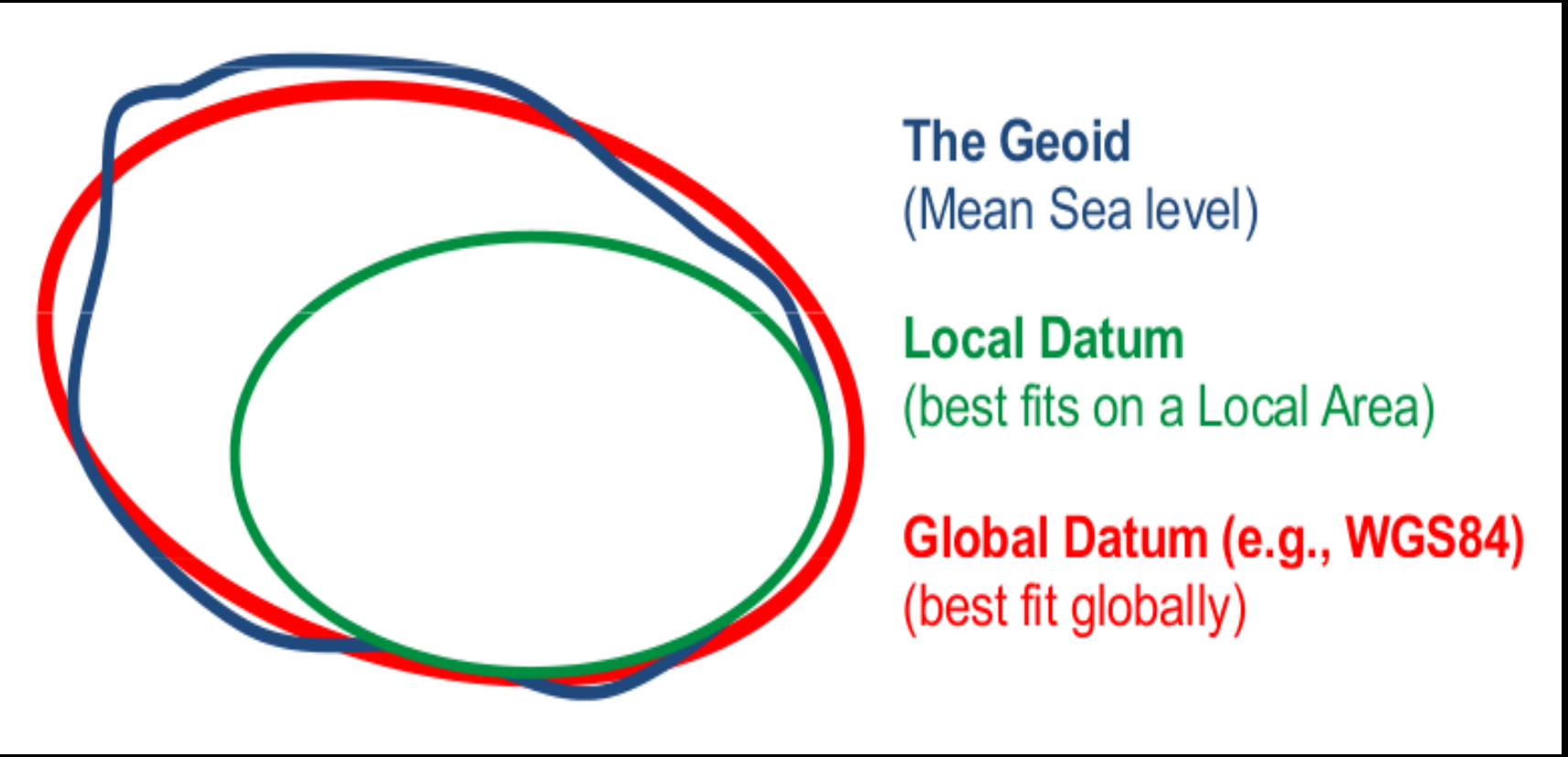
- Type of decimal point
- Order of longitude, latitude
- Degree minute second vs degree decimal minute
- Padding with 0

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

Geodetic datum

Geodetic datum

Most of the datasets we received lack this information



Geodetic datum

decimalLongitude	decimalLatitude	geodeticDatum
-67.83603	-67.30107	EPSG:4326

Use EPSG code of the spatial reference system

EPSG code for World Geodetic System 1984 (WGS 84) is EPSG:4326

Georeferencing best practice

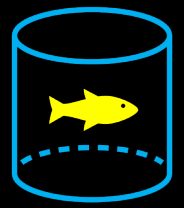
Georeferencing Best Practice from GBIF
http://mb.gbif.org/documents/doc-georeferencing-best-practices/en/

Filling in the occurrence data template

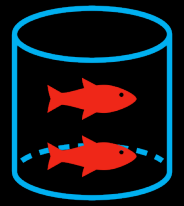
Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net → RMT Net 1, RMT Net 2



Jar 01



Jar 02

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientificName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326

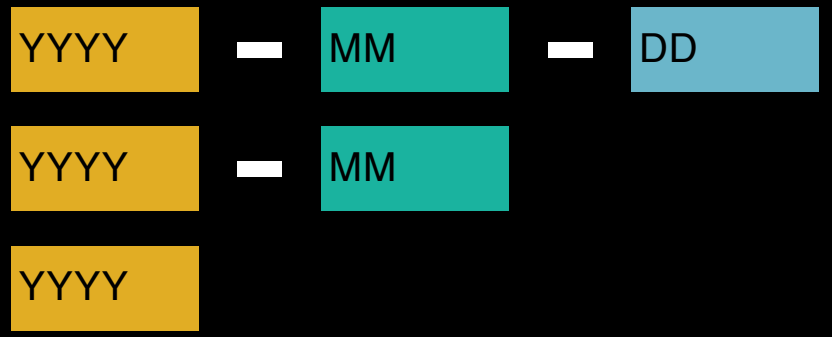
eventDate


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27½-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13
Hisssss
2-27-13

Date

ISO 8601 standard

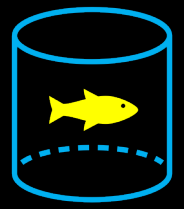
Fixed number of digits padded with leading zero



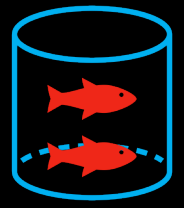
Filling in the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200



Jar 01



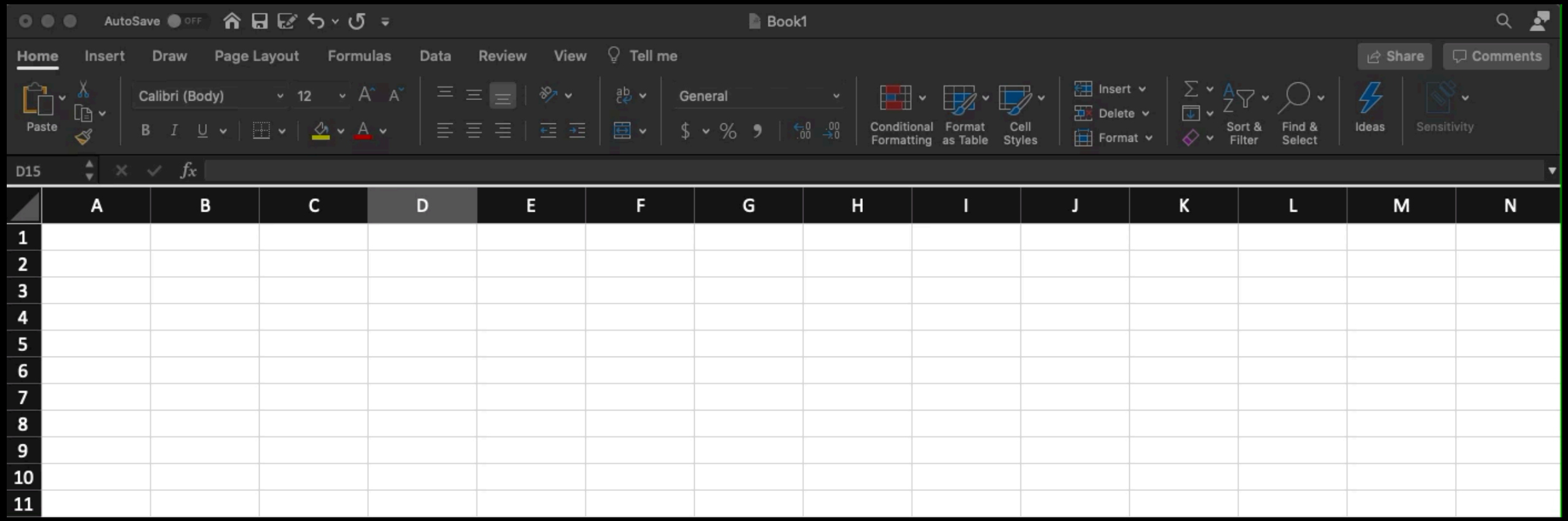
Jar 02

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientificName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02

Excel saves what you see, not what you typed

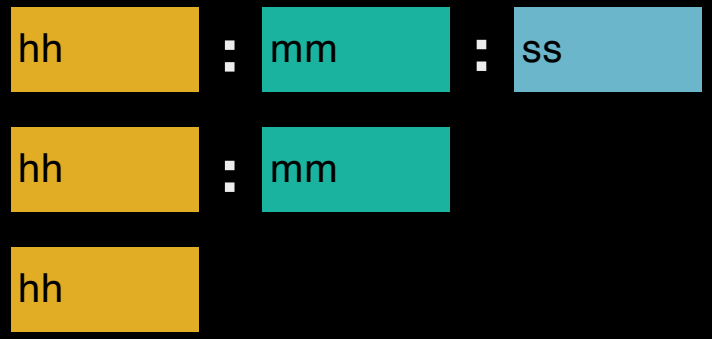


Microsoft Excel version 16.42 (2020)

Time

ISO 8601 standard

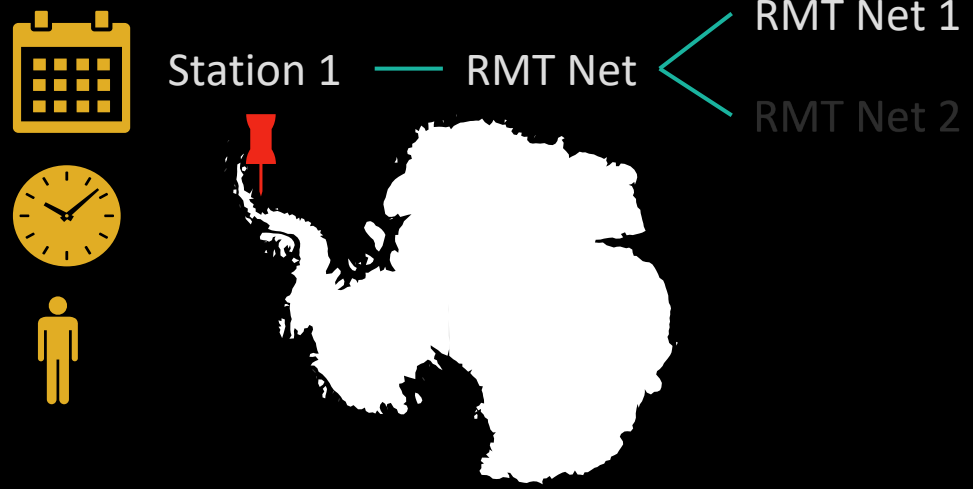
Fixed number of digits padded with leading zero



Filling in the data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200



Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientifiCName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day	eventTime
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00

Filling in the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net → RMT Net 1, RMT Net 2

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientifiCName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day	eventTime
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00

Time zone

Time zone is very often neglected in the datasets we received

Time zone

ISO 8601 standard

Z	UTC
±hh:mm	Offset from UTC

Time zone

ISO 8601 standard

Example	Description
14:07-06:00	2:07 pm in the time zone 6 hours earlier than UTC
08:40:21Z	8:40:21am UTC

Filling in the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net → RMT Net 1, RMT Net 2

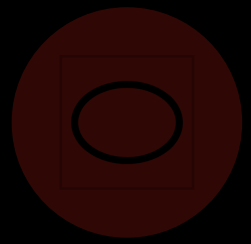
Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientifiCName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day	eventTime
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00

occurrenceID	timeZone
ANT1_Stn1_RMT-Net_1_Jar01	Z
ANT1_Stn1_RMT-Net_1_Jar02	Z

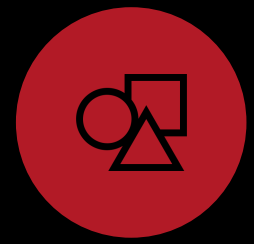
WHAT



WHERE

WHEN

WHO



HOW

Filling in the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net → RMT Net 1, RMT Net 2

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientifiCName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day	eventTime
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG4326	2020-11-02	2020	11	02	08:32:00

occurrenceID	timeZone	identifiedBy	eventID	samplingProtocol	minimumDepthInMeters	maximumDepthInMeters
ANT1_Stn1_RMT-Net_1_Jar01	Z	John Doe	ANT1_Stn1_RMT-Net_1	RMT Net	500	1000
ANT1_Stn1_RMT-Net_1_Jar02	Z	John Doe	ANT1_Stn1_RMT-Net_1	RMT Net	500	1000

Filling in the occurrence data template

Field event sheet
eventID: RMT Net

Sampling events		
event	start depth	end depth
	m	m
RMT Net 1	1000	500
RMT Net 2	500	200

Station 1 — RMT Net → RMT Net 1, RMT Net 2

Field occurrence sheet
eventID: RMT Net 1

Sample ID	Species	Count
Jar 01		1
Jar 02		2

occurrenceID	scientificName	individualCount	basisOfRecord	decimalLongitude	decimalLatitude	geodeticDatum	eventDate	year	month	day	eventTime
ANT1_Stn1_RMT-Net_1_Jar01		1	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00
ANT1_Stn1_RMT-Net_1_Jar02		2	PreservedSpecimen	-67.83603	-67.30107	EPSG:4326	2020-11-02	2020	11	02	08:32:00

occurrenceID	timeZone	identifiedBy	eventID	minimumDepthInMeters	maximumDepthInMeters	measurement
						unit
ANT1_Stn1_RMT-Net_1_Jar01	Z	John Doe	ANT1_Stn1_RMT-Net_1	500	1000	Value 1
ANT1_Stn1_RMT-Net_1_Jar02	Z	John Doe	ANT1_Stn1_RMT-Net_1	500	1000	Value 2

SAMPLING EVENT | REGISTERED MAY 26, 2020

Epipelagic mesozooplankton distribution and abundance in Southern Ocean Atlantic sector and the North Atlantic and Arctic 1996-2013

Published by British Antarctic Survey

Peter Ward • Geraint Tarling • Rachael Shreeve • Petra ten Hoopen

Event data

samplingProtocol	samplingEffort	sampleSizeValue	sampleSizeUnit
https://www.bodc.ac.uk/resources/inventories/cruise_inventory/report/5386/	average of 30 min net deployment	1	square metre

Project metadata

Methodology

Study extent

Data were gathered on a series of oceanographic cruises aboard the RRS James Clark Ross during expeditions to the Atlantic Sector of the Southern Ocean and the North Atlantic/Arctic.

Sampling

Samples were retrieved by either a motion compensation Bongo net or a mini Bongo net, and then preserved with Borax buffered 10% formalin for analysis back at the home laboratory. Taxa were identified through examination by light microscopy. The abundance of each taxa within a sample was determined through examining a known fraction of the sample and then making an inverse multiplication of that fraction. Known fractions were principally achieved through the use of a Folsom splitter. Volumetric abundances (individuals m-2) were determined through dividing sample abundance by the volume of water sampled by the respective net. Net volume was mainly derived through multiplying the net opening diameter by the maximum depth of sampling. This calculation assumes 100% net sampling efficiency and that specimens were only captured during the ascent phase of the net deployment.

Method steps

All species were identified according to the taxonomic guides available at time of analysis and the user must be aware that some species names may have since been updated.

https://www.gbif.org/dataset/402c81b6-8d69-4650-8826-92f91d2728a7

Summary

- Use original identifiers
- Basis of record
- Record the most certain taxonomy
- Occurrence status: present/absent
- Atomize columns for latitude and longitude – record what you see, transform later
- Take precision into account
- Geodetic datum
- Atomize columns for date
- Time and time zone
- Sampling protocol and environmental measurement

Citations and resources

- Cartoons by XKCD more at https://m.xkcd.com/
- www.biodiversity.aq
- www.obis.org
- www.gbif.org
- https://scar.github.io/EGABIcourse19/metadata-darwin-core.html
- https://www.scar.org/science/egabi/
- Georeferencing Best Practice from GBIF http://mb.gbif.org/documents/doc-georeferencing-best-practices/en/

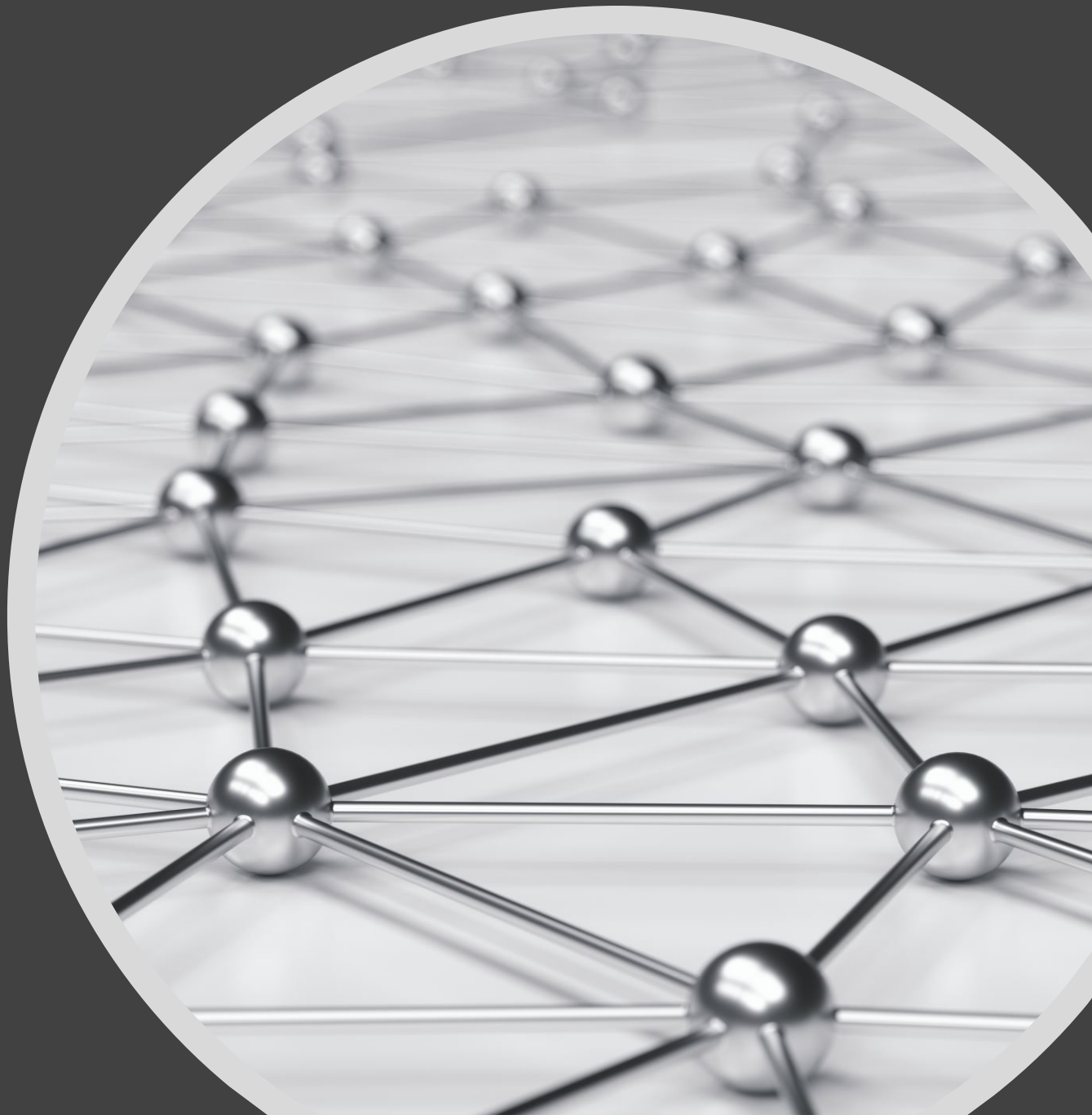
Acknowledgement

- Scientists who contributed to templates design:
  Huw Griffiths, Anton Van de Putte, Fokje Schaafsma, Hauk Flores

- Early career scientists who provided valuable input:
  Robyn Samuel, Raissa Meyer, Louraine Salabao, Jasmine Lee, Kimberlee Bradley, Svenja Hafter

Thank you!

Additional resources



Templates

- https://sios-svalbard.org/cgi-bin/darwinsheet/index.cgi?setup=darwin

Taxonomic data

- World Register of Marine Species (WoRMS) initial focus on marine species but extending to terrestrial and species traits. Also Register of Antarctic species
- Integrated Taxonomic Information System(ITIS) authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world.
- Species 2000 is an autonomous federation of taxonomic database custodians, involving taxonomists throughout the world
- Catalogue of Life is the most comprehensive and authoritative global index of species currently available. It consists of a single integrated species checklist and taxonomic hierarchy.

Genomic data

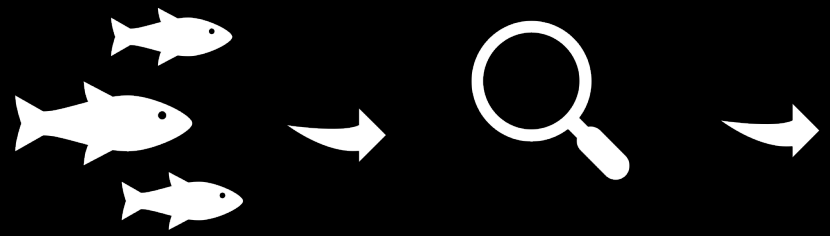
- The Genomic Standards Consortium (GSC) aims to make genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards

- Global Genome Biodiversity Network (GGBN) network of well-managed collections of genomic tissue samples from across the Tree of Life, benefiting society through biodiversity research, development and conservation. This network will foster collaborations among repositories of molecular biodiversity in order to ensure quality standards, improve best practices, secure interoperability, and harmonize exchange of material in accordance with national and international legislation and conventions.

Example: genomics data

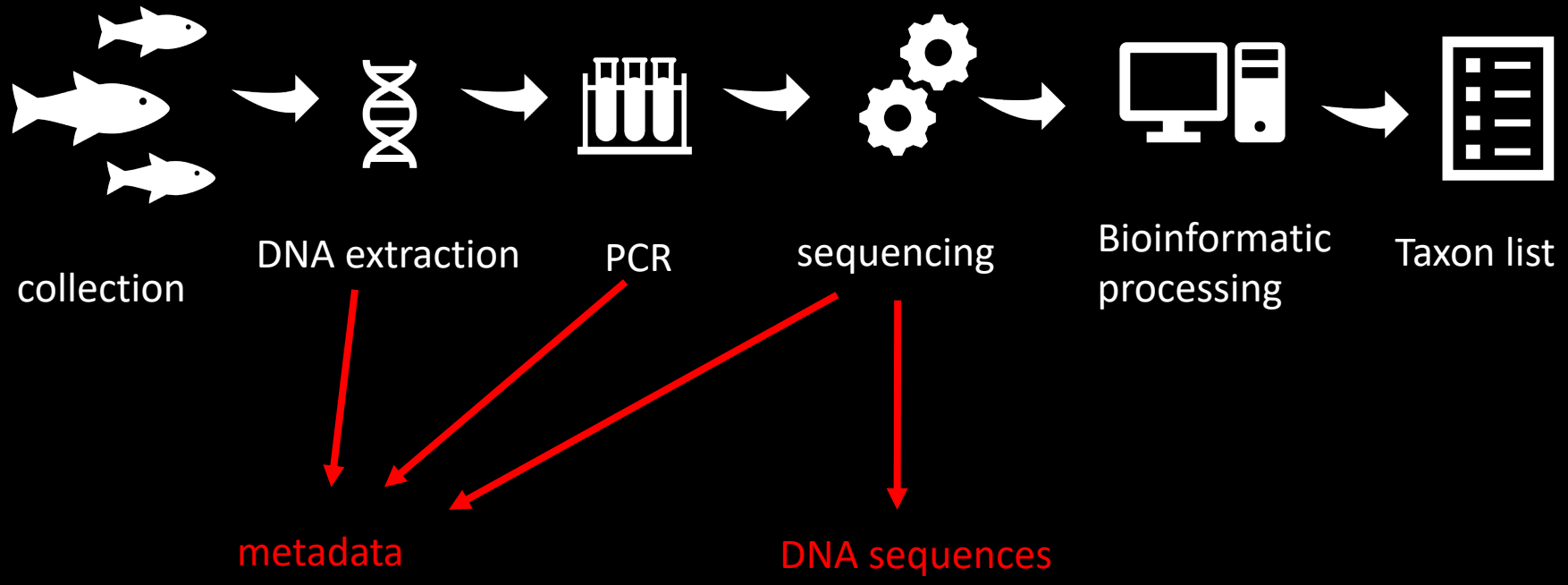
Genomic data is complex

occurrence

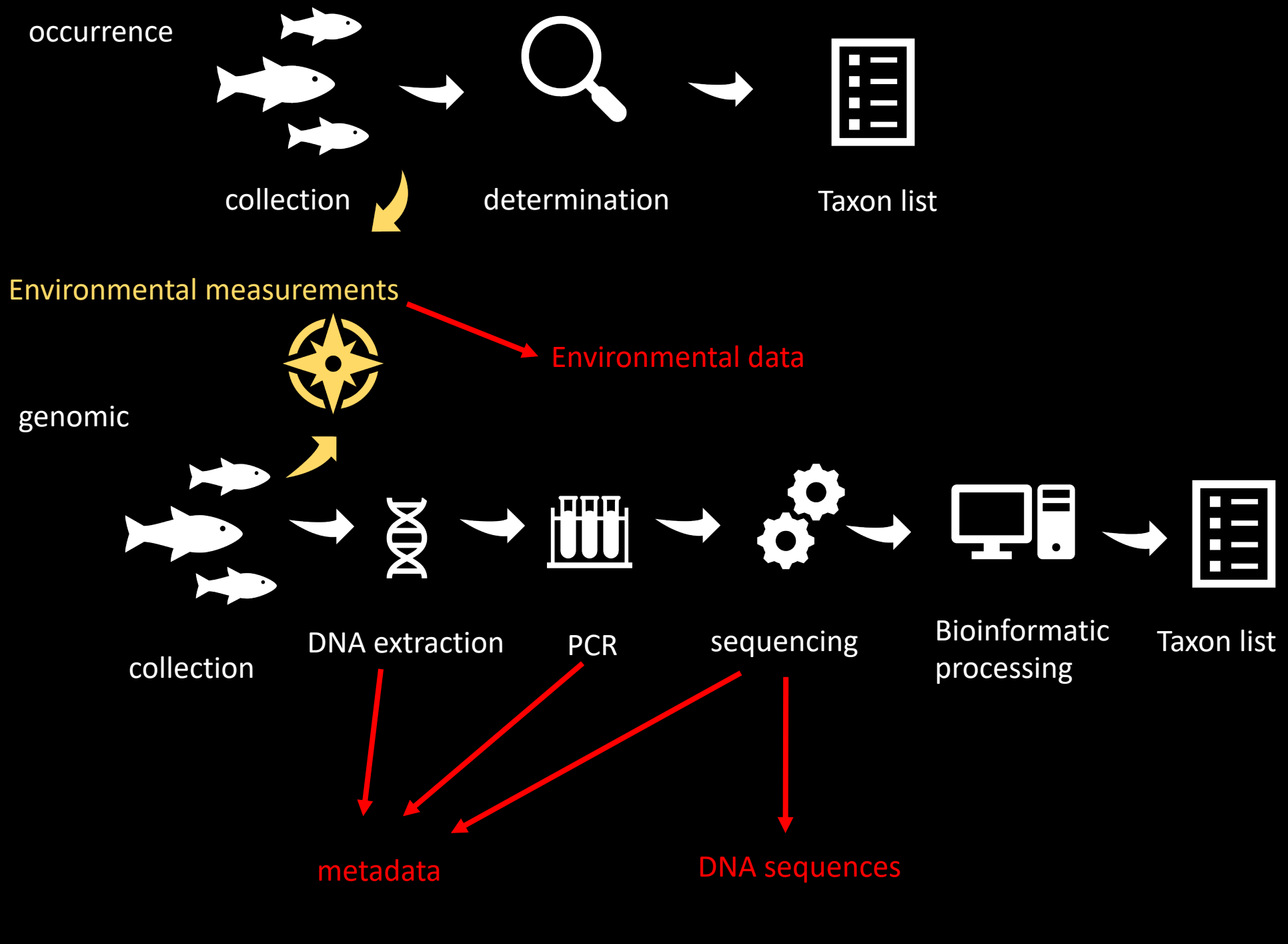


collection determination Taxon list

genomic



Genomic data is complex



Metadata and environmental data

- Every action that can affect the data
  - Sample collection, culturing or experimental settings
  - DNA extraction and sequencing protocols
  - essential to interpret the sequence data

- Associated Environmental measurements
  - Date + geographic coordinates
  - Any other measurement taken at the time
    - E.g. pH, ions, conductivity, weather conditions,...

- Put bluntly: sequence data is useless without

MIxS standard

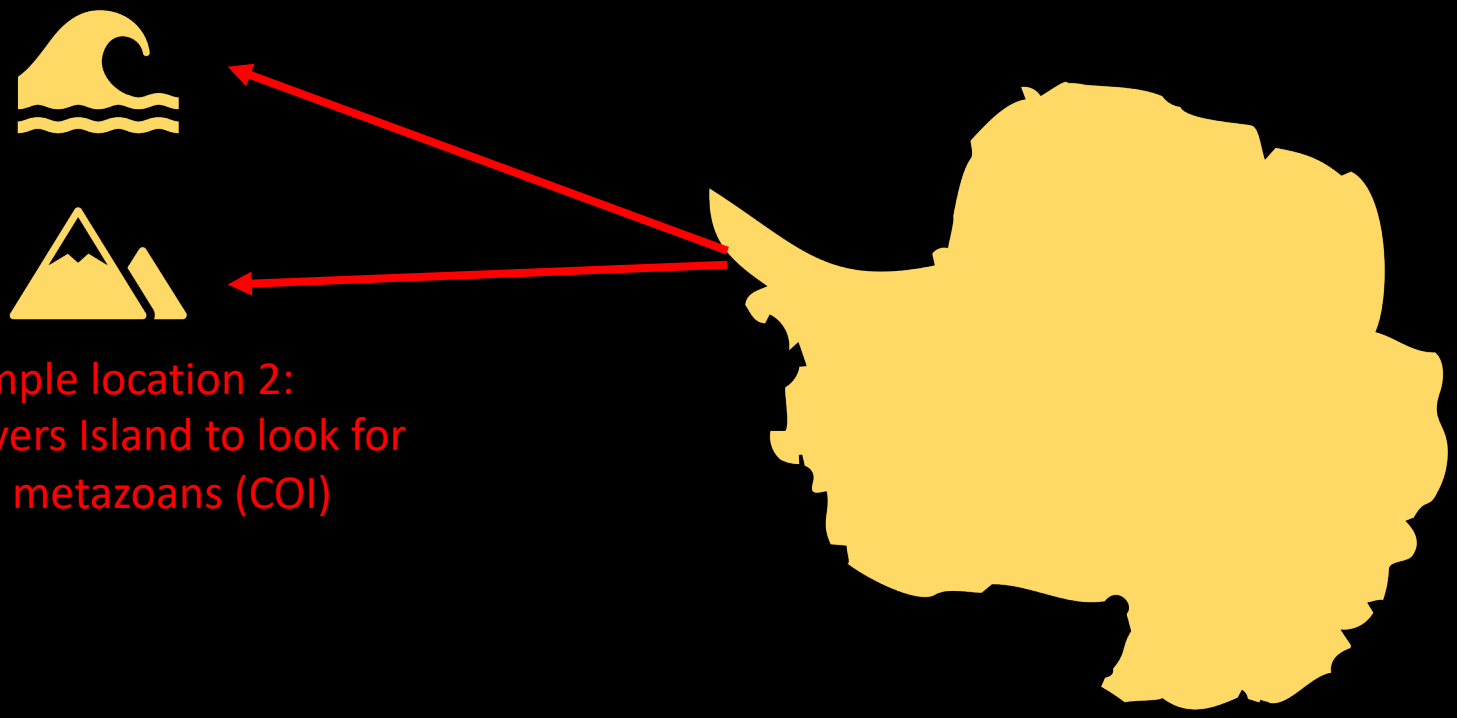
- Developed by: Genomics Standards Consortium (GSC)
- Minimum Information on Any Sequence (MIxS)
  - a list of standardized terms
    = a minimum set to describe sequences
    + any other relevant information
  - Allows standardization of environmental measurements
  - Accepted (required) by ENA and NCBI

List of the MIxS terms → https://gensc.org/mixs/

MIxS example

Example: 2 samples on 1 sampling day:

Sample location 1: Gerlache strait seawater for 16S rRNA



Sample location 2:
Anvers Island to look for
soil metazoans (COI)

MIxS example

	MIxS_Term_1	MIxS_Term_2	MIxS_Term_3				
1 	value						
2 							

Rows = sample = corresponds to an event

Columns = variable = standard MIxS terms

MIxS example

Cf. DawinCore eventID

Cf. DawinCore parentEventID

	sample_name	project_name					
	2020_11_Gerlache_day1s1	webinaR2020					
	2020_11_Anvers_day1_s2	webinaR2020					

- Sample_name:
  - choose something logical (e.g. place_date_letter_number)
  - Only underscore or capitals, no special characters!

MIxS example

sample_name	project_name	**lat_lon**	decimalLatitude	decimalLongitude			
2020_11_Gerlache_day1s1	webinar2020	-67.43 123.09	-67.43	123.09			
2020_11_Anvers_day1_s2	webinar2020	-69.43 123.59	-69.43	123.59			

- lat_lon
  - Space separated
  - Decimal degrees
  - Write lat and lon in separate fields (e.g. “decimalLatitude”, “decimalLongitude”)

MIxS: required terms

sample_name	project_name	decimalLatitude	decimalLongitude	**collection_date**	year	month	day
2020_11_Gerlache_day1s1	webinar2020	-67.43	123.09	2020-11-06	2020	11	06
2020_11_Anvers_day1_s2	webinar2020	-69.43	123.59	2020-11-06	2020	11	06

- collection_date
  - Use ISO format (YYYY-MM-DD)
  - Best also include year, month, day

MIxS: required terms

sample_name	project_name	**geo_loc_name**	**seq_meth**	**env_biome**	**env_feature**		
2020_11_Gerlache_day1s1	webinar2020	Gerlache Stait	Illumina MiSeq 300bp-PE	Ocean ENVO:01000048	Marine pelagic ENVO_01000044		
2020_11_Anvers_day1_s2	webinar2020	Anvers Island	Illumina MiSeq 300bp-PE	Polar ENVO:01000339	Glacial feature ENVO_00000131		

- Free text fields: try to standardize if possible
  - Use a ontologies (e.g. ENVO: http://bioportal.bioontology.org/ontologies/ENVO)

MIxS: environmental data

- The environmental package
  - Different packages of terms specific to one environment
  - E.g. soil, water, host_associated,...

  - To provide some structure in long list of terms
  - Focused on associated measurements
    - soil: depth, pH,...
    - water: cond, turbidity, part_org_nitro, phosphate,...

  - Pick and mix, as long as you use as many standardized terms as possible

MIxS: environmental data

sample_name	**depth**	part_org_nitro					
units	meter	micromole per liter					
2020_11_Gerlache_day1s1	100	0.4					
2020_11_Anvers_day1_s2	0	NA					

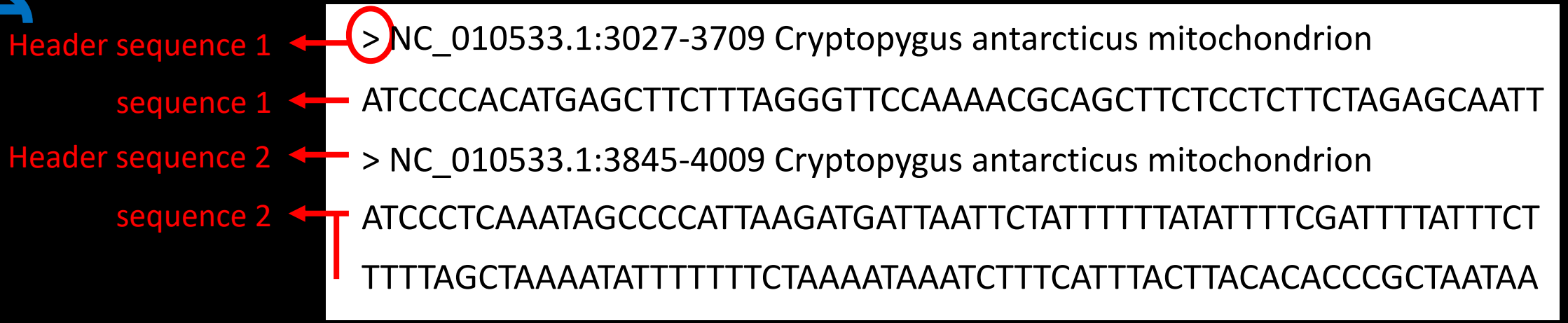
- Include units, and write in full
- Also include any information that has no appropriate MIxS term
- New terms can be requested/proposed to GCS though their GitHub issues
- Only use “NA” where there is no information (not: “ND”, “NULL”, “0”,...)

MIxS: sequencing terms

- Best also include
  - pcr_primers
  - target_gene; target_subfragment
  - subspecf_gen_lin
    - = target taxonomic group
  - samp_collect_device and samp_mat_process
  - nucl_acid_amp; nucl_acid_ext; pcr_cond,...

DNA and RNA sequences

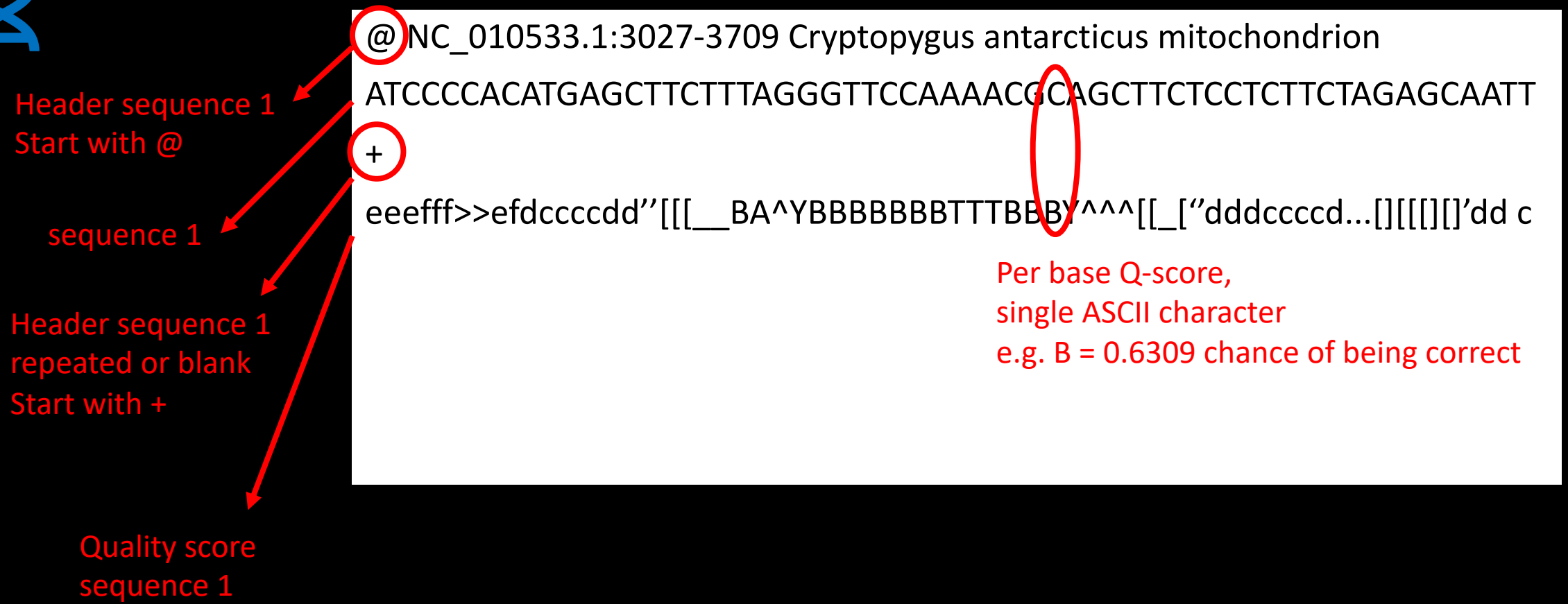
- FASTA format (=text file)



- IUPAC notation of bases
  - Known bases: A, T, G, C
  - Ambiguous bases: W (=A or T), S (=C or G), R, Y, K, M, B, D, H, V, N

DNA and RNA sequences

- FASTQ variant (=FASTA + quality score)



MIxS useful terms

sample_name	**pcr_primers**	**target_gene**	**subspecf_gen_lin**	nucl_acid_ext			
2020_11_Gerlache_day1s1	ATTGTGAAGTAAGGTGTCW	16S rRNA	Bacteria Archaea	dx.doi.org/10.17504/protocols.io.bdwsi7ee			
2020_11_Anvers_day1_s2	CCTGHAATCGGTTGGAGCA	COI	Metazoans	dx.doi.org/10.17504/protocols.io.bdwsi7ee			

- Don’t use “, “ as separator (cf. CSV file)
- For protocols: use DOIs or URLs to the publications

Using MIxS

- Possible to upload alongside sequences to NCBI or ENA
- Can be added to a GBIF DarwinCore archive
  - Using event core (=list of sequence samples)
  - Add MIxS file as extended Measurement or Fact (eMoF) extension
  - See https://docs.gbif-uat.org/publishing-dna-derived-data/1.0/en/#environmental-dna-as-a-source-for-dna-derived-occurrence-data
- List the dataset on POLA3R
  - Specific data repository of any Antarctic sequence data
  - internal database (searchable metadata)
  - link to sequences on INSDC (not in MIxS)
  - www.biodiversity.aq/pola3r

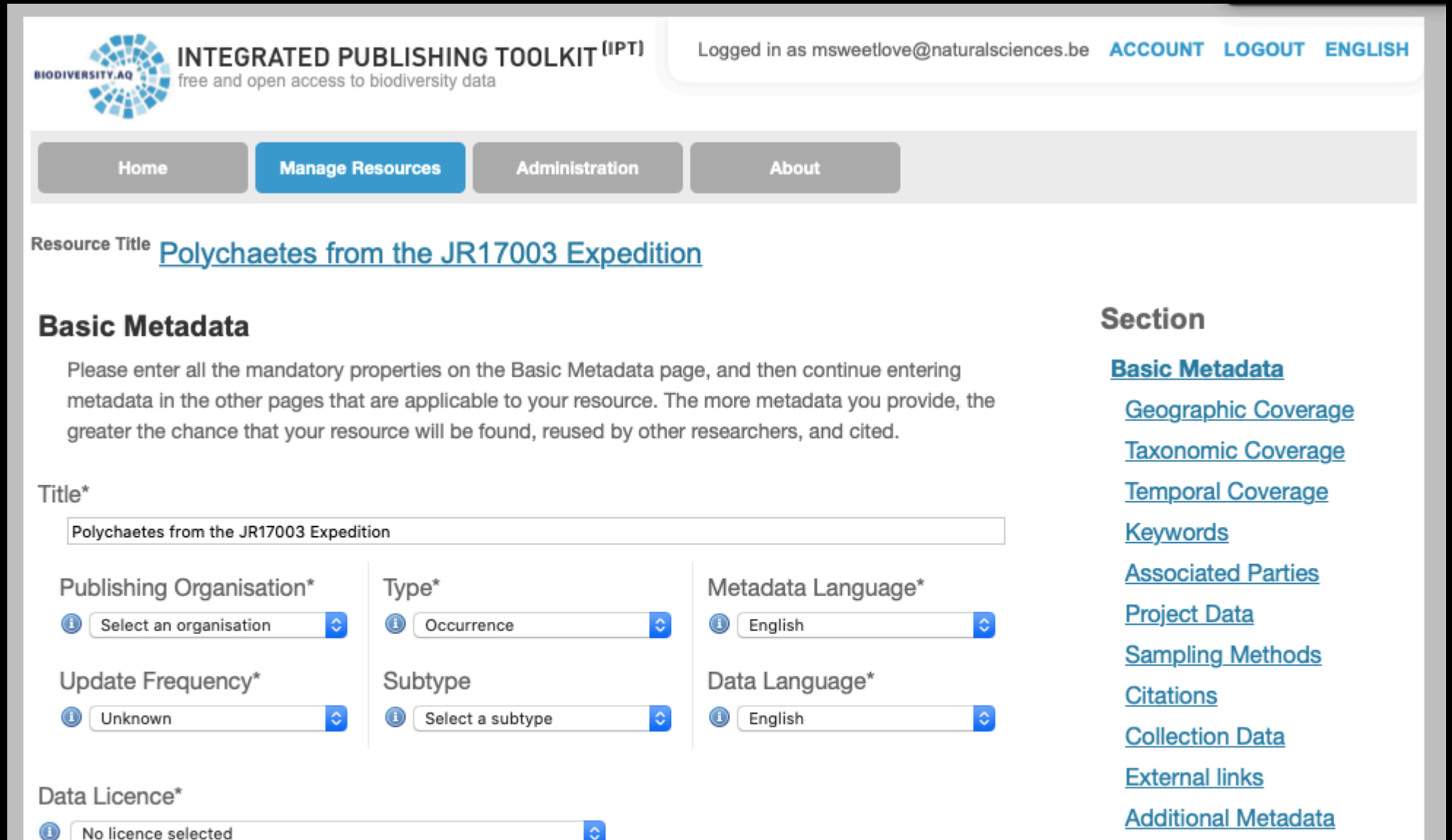
Thank You

way to deal with metadata

1. In the publication
   - Not physically linked with the sequence data
   - Often to concise to be fully reproducible
   - No standardization
   + High visibility in the scientific community
2. Register to a biodiversity repository (e.g. GBIF or POLA3R)
   + Better linkage to the sequence data
   + May include environmental metadata
   + Possibility to be very extensive
   + Standardized format (EML=Ecological Metadata Language and/or MIxS)
3. Attach to the sequence data on ENA or NCBI
   - But difficult to find or download
   + Physically linked with the sequence data
   + Standardized format (MIxS)
   + Ideal for PRE-sampling

Register metadata a biodiversity repository

https://ipt.biodiversity.aq



Resource Title Polychaetes from the JR17003 Expedition

Sampling Methods

Please enter metadata about the sampling methods used for the data represented by the resource.

Study Extent*

Sampling Description*

Quality Control

Step Description*

Add new method step

This XML file does not appear to have any style information associated with it. The document tree is shown below.

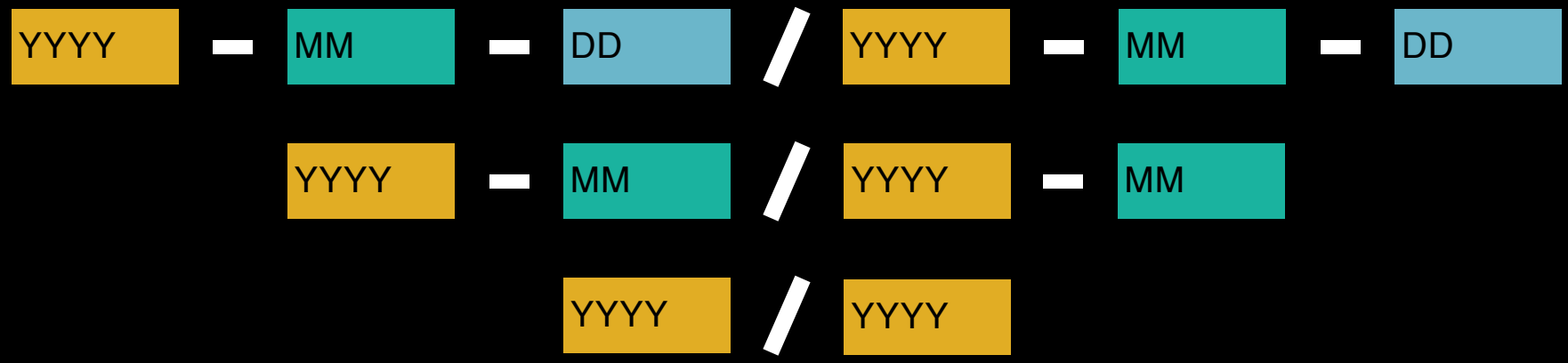

```
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.1" xmlns:dc="http://purl.org/dc/terms/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 http://rs.gbif.org/schema/eml-gbif-profile/1.1/eml.xsd" packageId="https://ipt.biodiversity.aq/resource?
 id=2503_antarctic_nemertea/v1.0" system="http://gbif.org" scope="system" xml:lang="eng">
  <dataset>
    <alternateIdentifier>
      https://ipt.biodiversity.aq/resource?r=2503_antarctic_nemertea
    </alternateIdentifier>
    <title xml:lang="eng">A collection of Antarctic Nemertea</title>
    <creator>
      <individualName>
        <givenName>Claude</givenName>
        <surName>De Boyer</surName>
      </individualName>
      <organizationName>Belgian Royal Institute of Natural Sciences</organizationName>
      <address>
        <city>Brussels</city>
        <country>BE</country>
      </address>
    </creator>
    <metadataProvider>
      <individualName>
        <givenName>Maxime</givenName>
        <surName>Sweetlove</surName>
      </individualName>
      <organizationName>Belgian Royal Institute of Natural Sciences</organizationName>
      <address>
        <city>Brussels</city>
        <country>BE</country>
      </address>
      <electronicMailAddress>msweetlove@naturalsciences.be</electronicMailAddress>
    </metadataProvider>
    <associatedParty>
      <individualName>
        <givenName>Claude</givenName>
        <surName>De Boyer</surName>
      </individualName>
      <organizationName>Belgian Royal Institute of Natural Sciences</organizationName>
      <address>
        <city>Brussels</city>
        <country>BE</country>
      </address>
      <role>user</role>
    </associatedParty>
    <pubDate>2020-09-10</pubDate>
    <langu        language>
    <abstrac
```


Safari

Date range

ISO 8601 standard

Start date / End date



etc etc …

Date, time intervals

ISO 8601 standard

2007/2008
2007-11/12
2007-11-13/15
2007-03-01/2008-05-11
2007-12-14T13:30/15:30Z
2007-03-01T13:00:00Z/2008-05-11T15:30:00Z