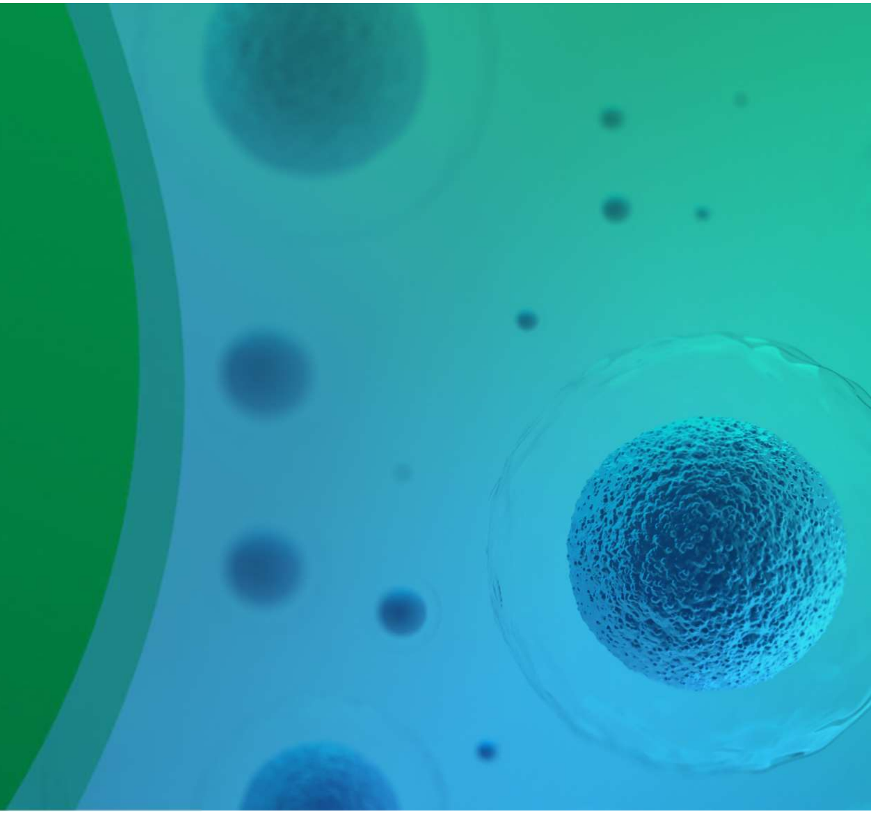
Data life cycle for Life Science



Stéphane PESANT

Senior marine biocurator

SZN Seminar, Naples, 2023-02-09

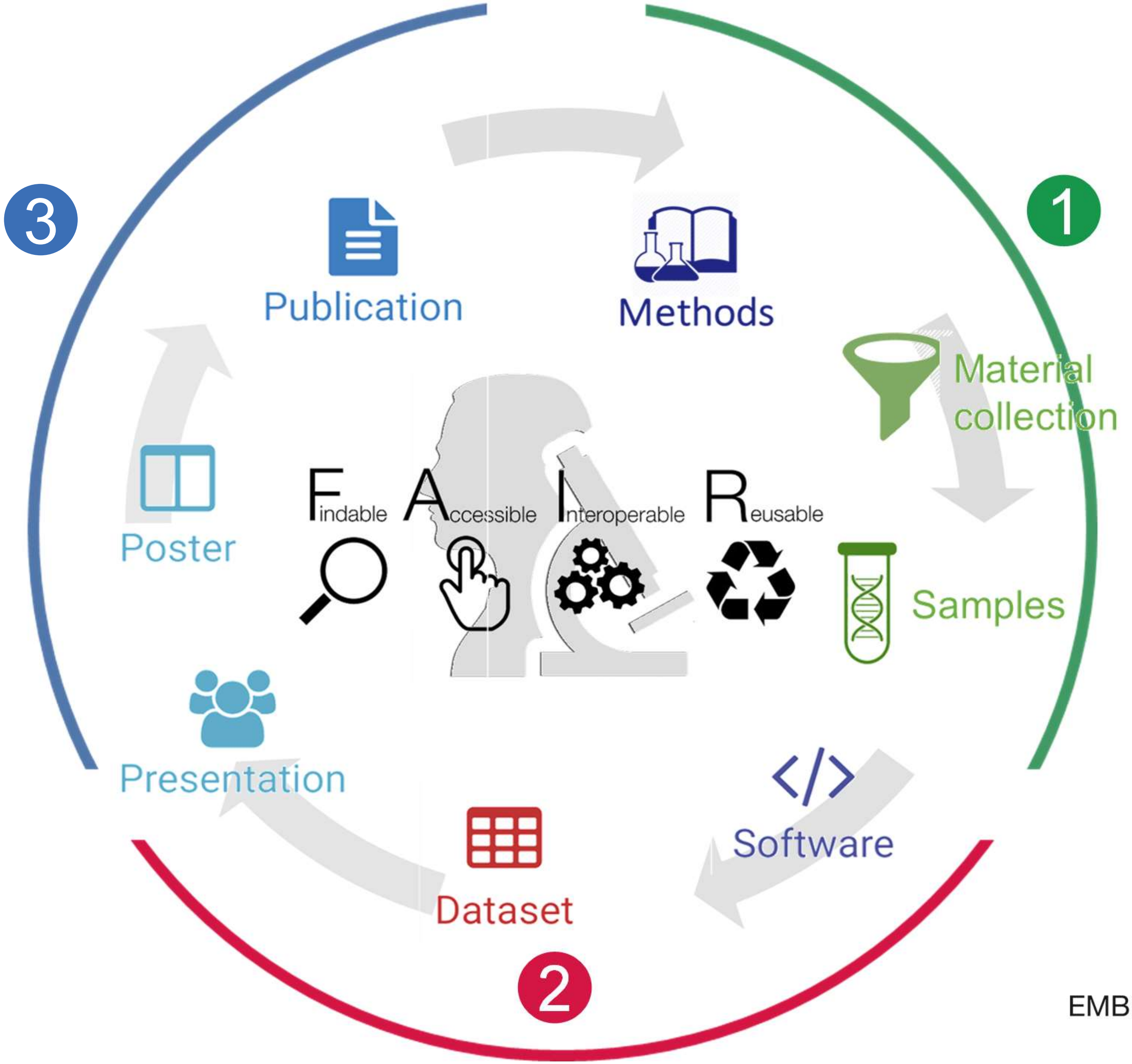
EMBL-EBI

Data life cycle

Sampling
Best Practices

Data Sharing
Best Practices

Publication
Best Practices



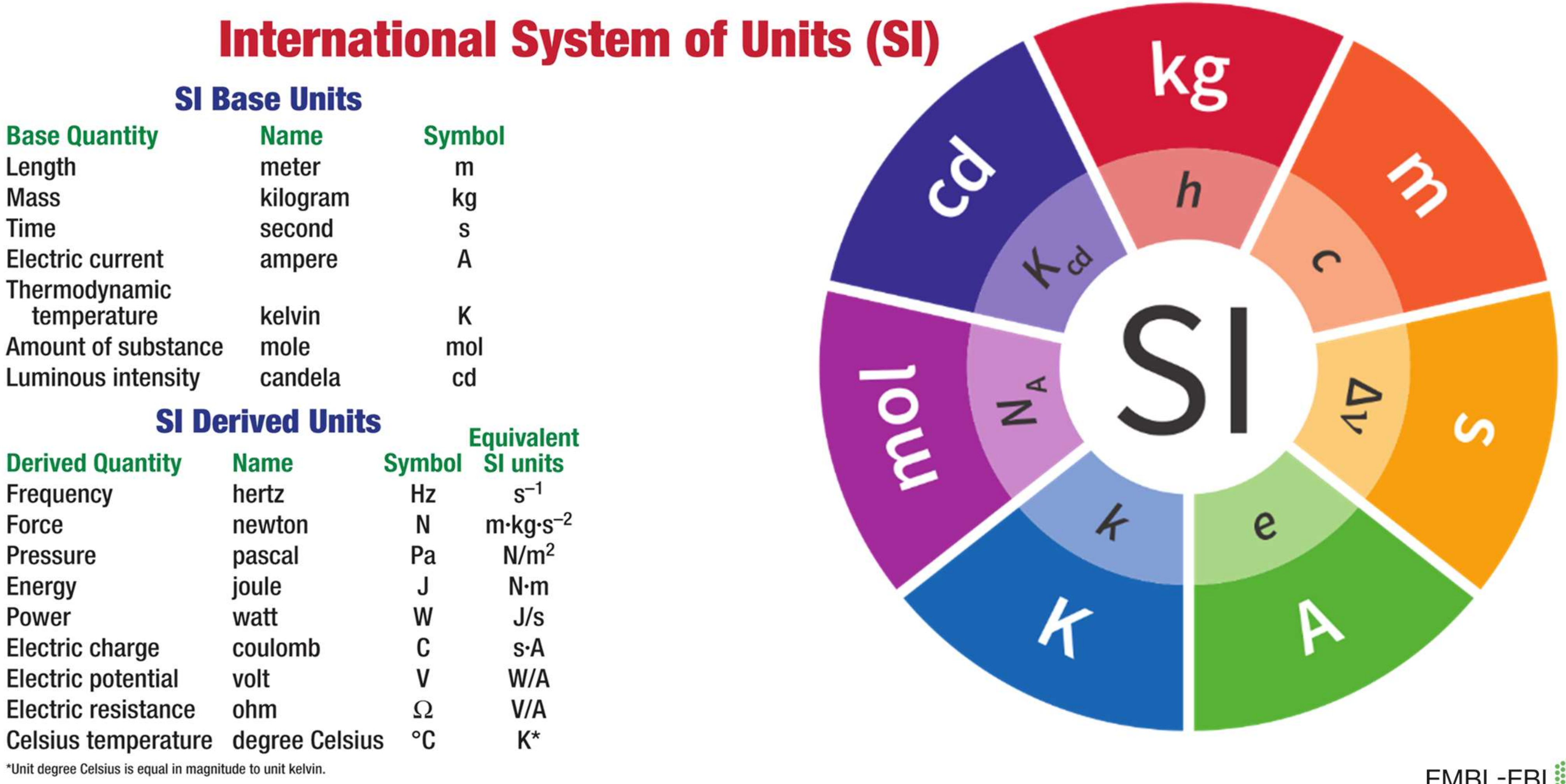
EMBL-EBI

Sampling best practices

Examplar initiatives - developing and experimenting with best practices

- Tara Oceans
- MicroB3
- EMO-BON
- AtlantECO

System of protocols



International System of Units (SI)

SI Base Units

Base Quantity	Name	Symbol
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

SI Derived Units

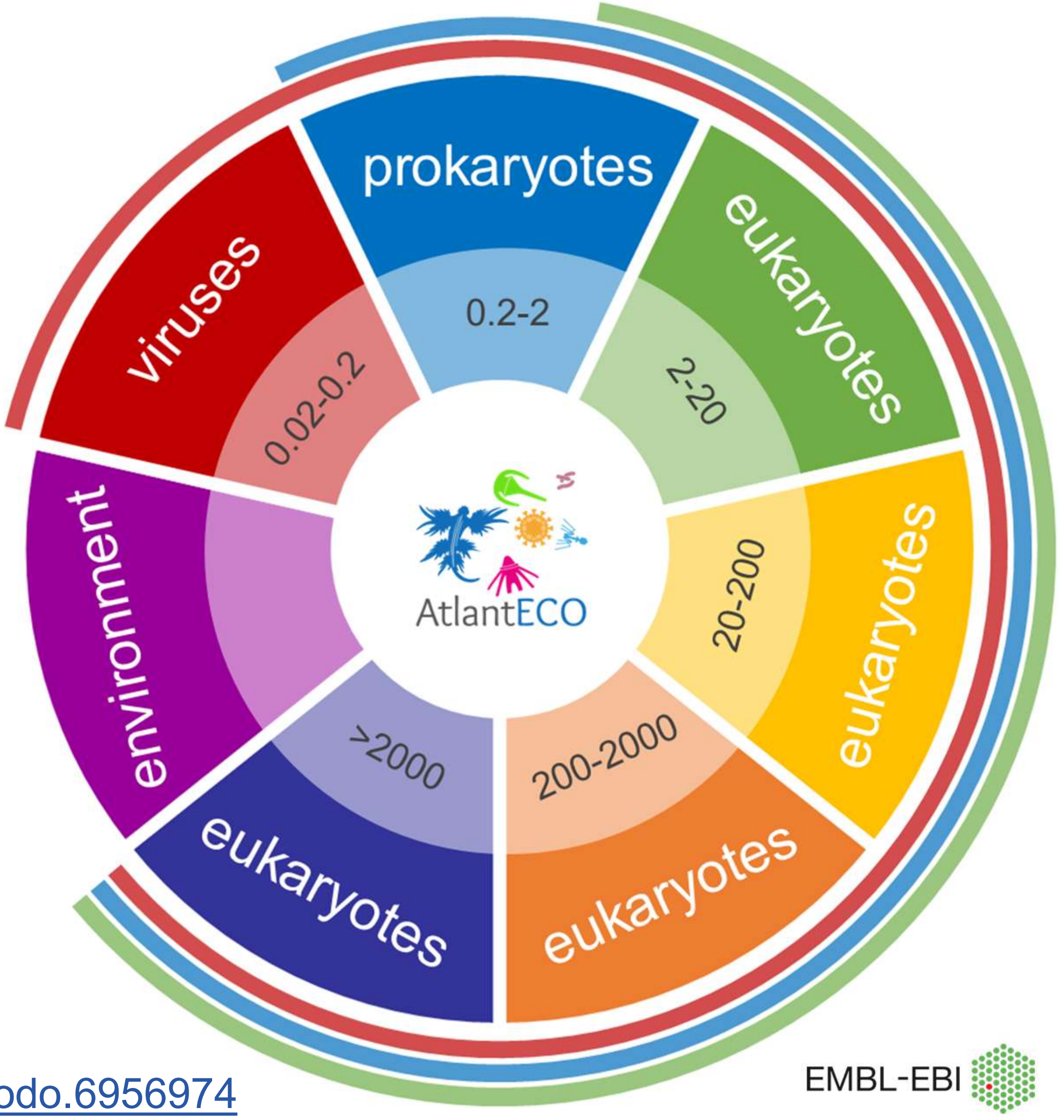
Derived Quantity	Name	Symbol	Equivalent SI units
Frequency	hertz	Hz	s^{-1}
Force	newton	N	$m{\cdot}kg{\cdot}s^{-2}$
Pressure	pascal	Pa	N/m^2
Energy	joule	J	N·m
Power	watt	W	J/s
Electric charge	coulomb	C	s·A
Electric potential	volt	V	W/A
Electric resistance	ohm	Ω	V/A
Celsius temperature	degree Celsius	°C	K*

*Unit degree Celsius is equal in magnitude to unit kelvin.

System of protocols

- Size
- Taxonomy
- Across size-fractions

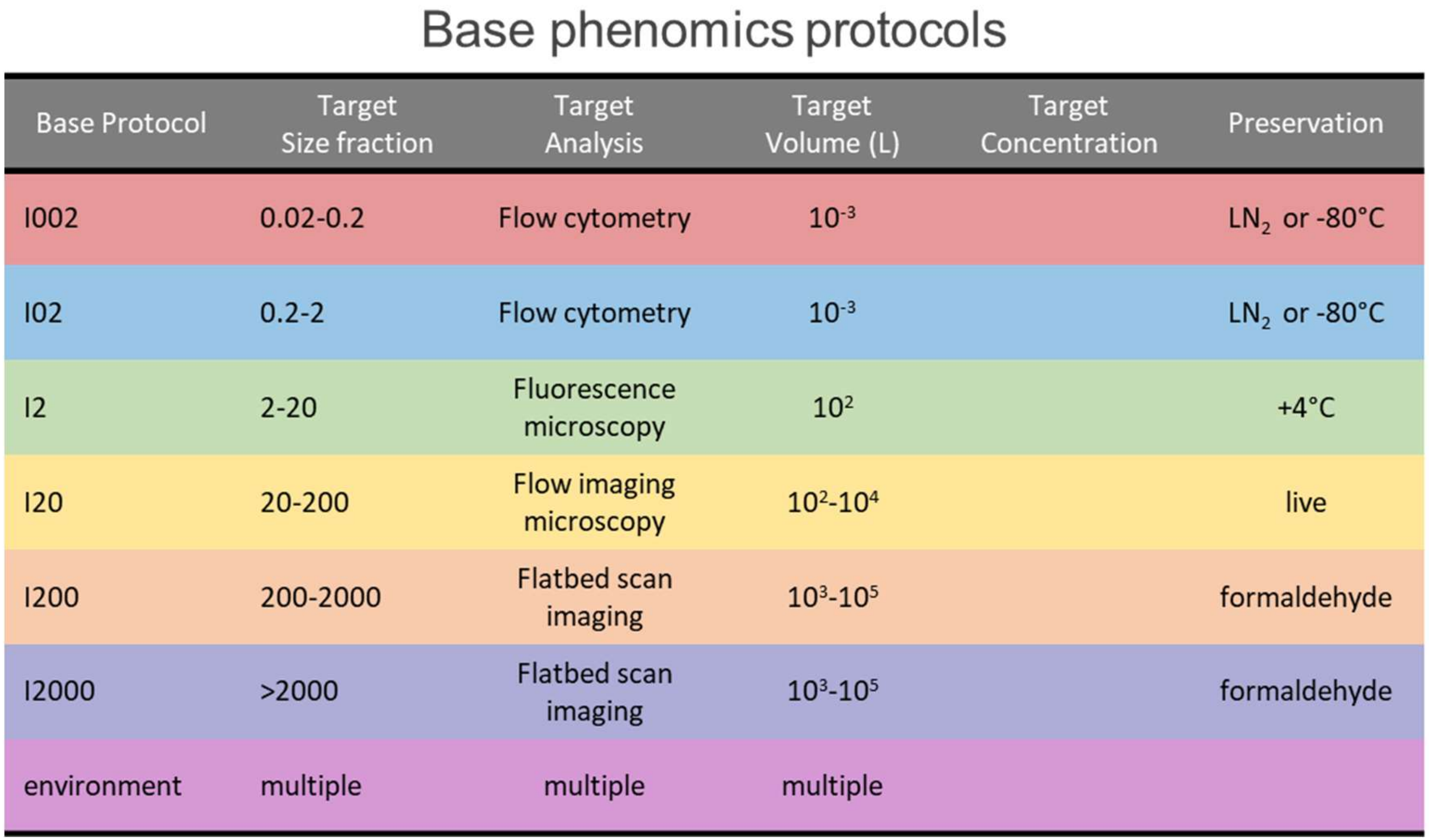
- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Phenomics



https://doi.org/10.5281/zenodo.6956974

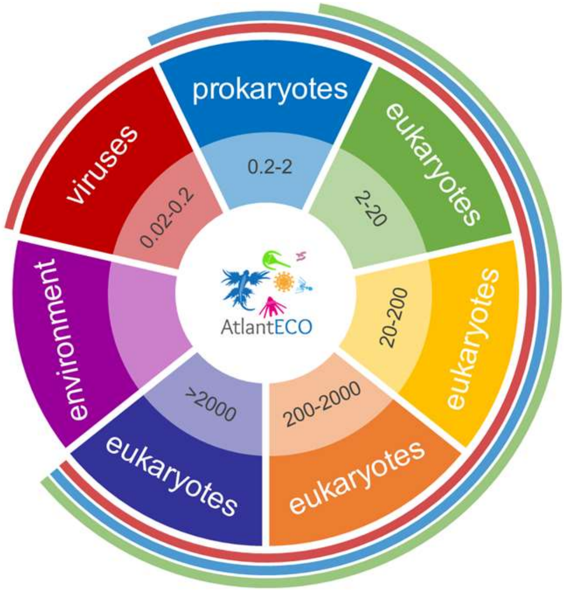
EMBL-EBI

System of protocols

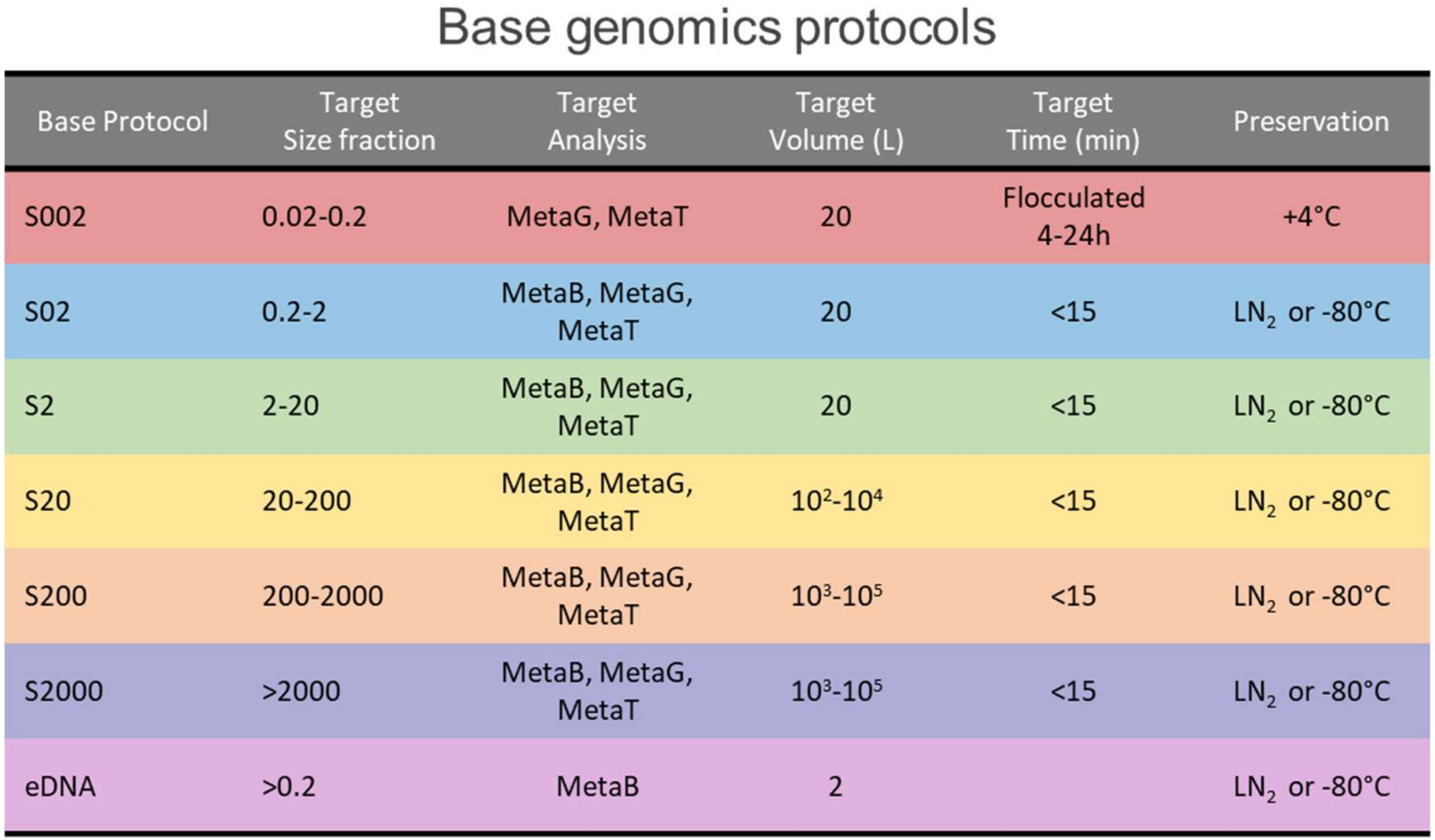


Base phenomics protocols

Base Protocol	Target Size fraction	Target Analysis	Target Volume (L)	Target Concentration	Preservation
I002	0.02-0.2	Flow cytometry	10^{-3}		LN$_2$ or -80°C
I02	0.2-2	Flow cytometry	10^{-3}		LN$_2$ or -80°C
I2	2-20	Fluorescence microscopy	10^{2}		+4°C
I20	20-200	Flow imaging microscopy	10^{2}-10^{4}		live
I200	200-2000	Flatbed scan imaging	10^{3}-10^{5}		formaldehyde
I2000	>2000	Flatbed scan imaging	10^{3}-10^{5}		formaldehyde
environment	multiple	multiple	multiple		

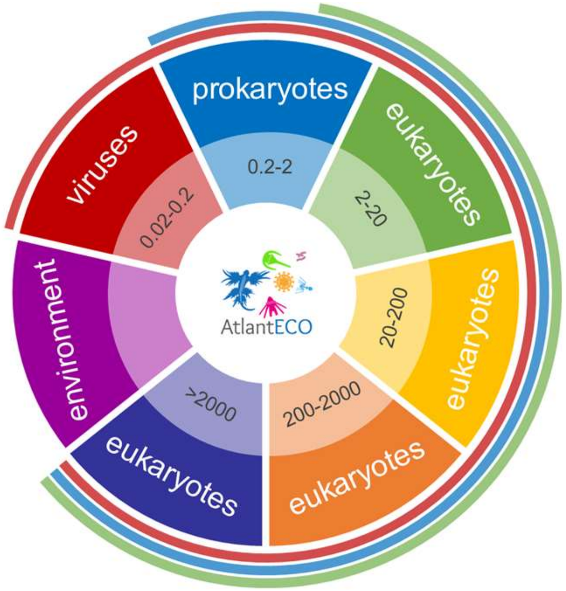


System of protocols



Base genomics protocols

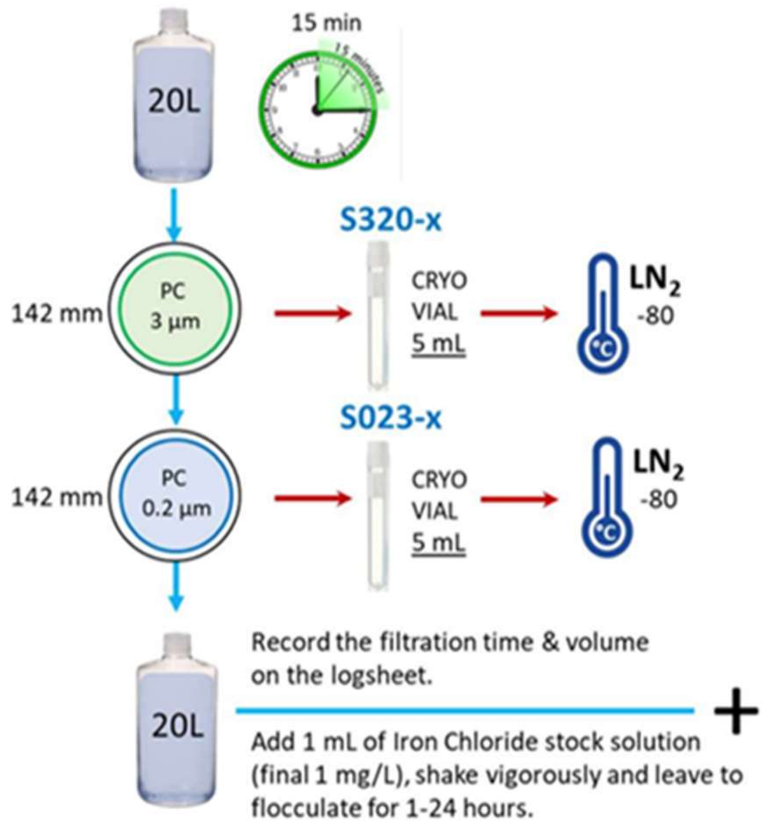
Base Protocol	Target Size fraction	Target Analysis	Target Volume (L)	Target Time (min)	Preservation
S002	0.02-0.2	MetaG, MetaT	20	Flocculated 4-24h	+4°C
S02	0.2-2	MetaB, MetaG, MetaT	20	<15	LN$_2$ or -80°C
S2	2-20	MetaB, MetaG, MetaT	20	<15	LN$_2$ or -80°C
S20	20-200	MetaB, MetaG, MetaT	10^2-10^4	<15	LN$_2$ or -80°C
S200	200-2000	MetaB, MetaG, MetaT	10^3-10^5	<15	LN$_2$ or -80°C
S2000	>2000	MetaB, MetaG, MetaT	10^3-10^5	<15	LN$_2$ or -80°C
eDNA	>0.2	MetaB	2		LN$_2$ or -80°C



Derived Protocols

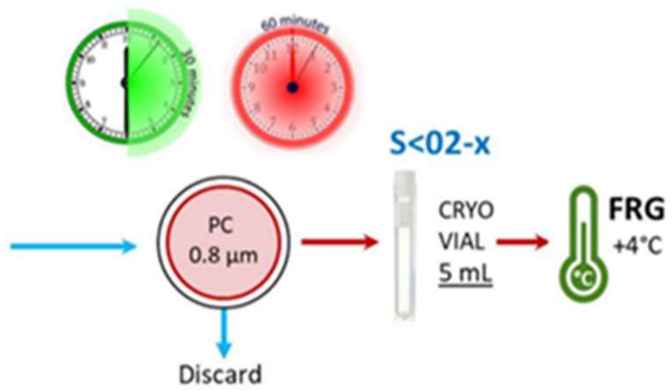
- Size fractions
- Filtration volume
- Filtration time
- Preservation method

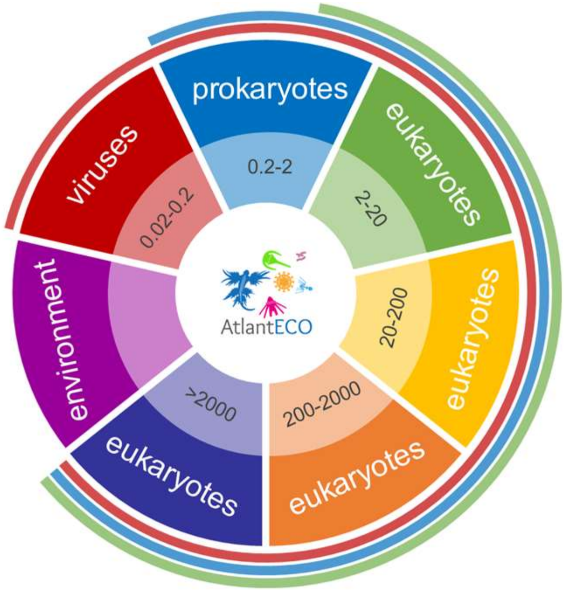
System of protocols



Record the filtration time & volume on the logsheet.

Add 1 mL of Iron Chloride stock solution (final 1 mg/L), shake vigorously and leave to flocculate for 1-24 hours.





Derived Protocols

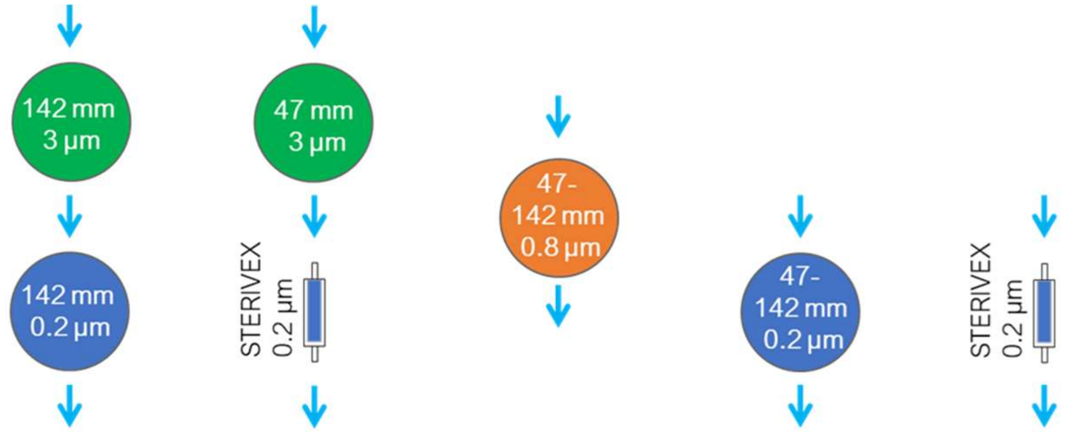
- Size fractions
- Filtration volume
- Filtration time
- Preservation method

Community Survey

All Atlantic Ocean Microbiome Sampling

OSD

Ocean Sampling Day



METHODS	A	B	C	D	E
Filters	2x 142 mm membranes	47 mm membr. + sterivex	47 or 142 mm membrane	47 or 142 mm membrane	sterivex
Vol. in 15 min.	10-20 L	2-5 L	1-10L	1-10L	1-10 L
Pump system	large peristaltic	small peristaltic	various	various	various
ADOPTED BY	**A**	**B**	**C**	**D**	**E**
Ocean Sampling Day			(✓)		✓
Bio-GO-SHIP*					✓
Tara Oceans	✓		(✓)	(✓)	
EMO-BON**	✓				

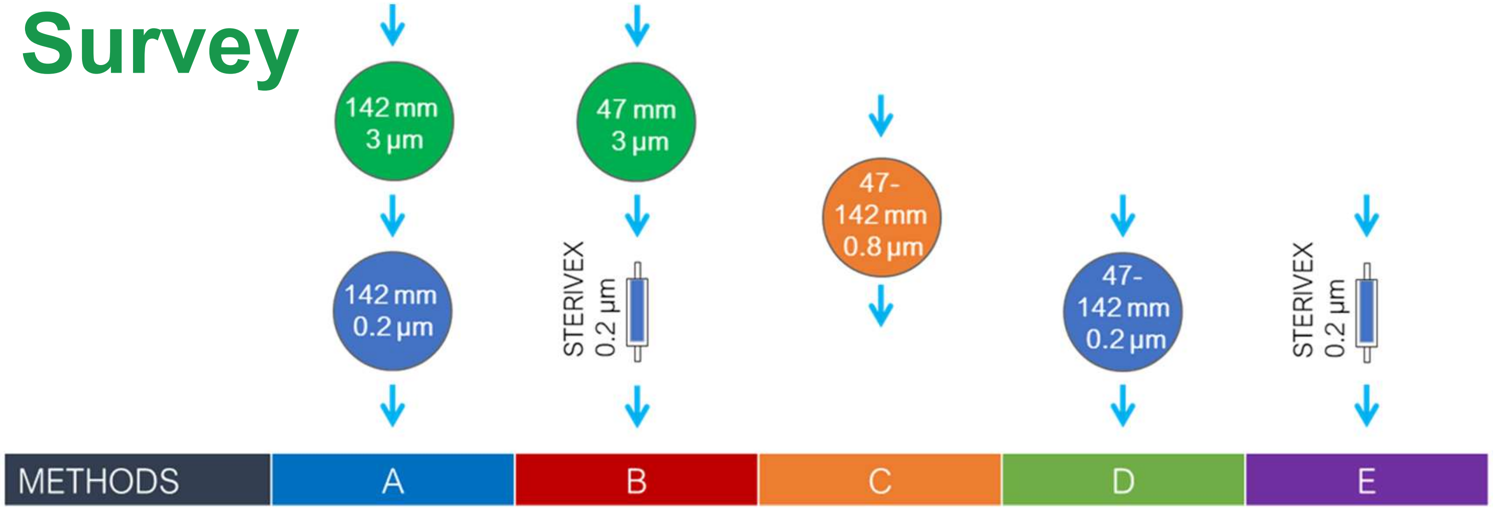
*Bio-GO-SHIP Linking marine biodiversity and biogeochemistry (https://biogoship.org/)
**European Marine Omics Biodiversity Observatory Network (https://www.embrc.eu/emo-bon)

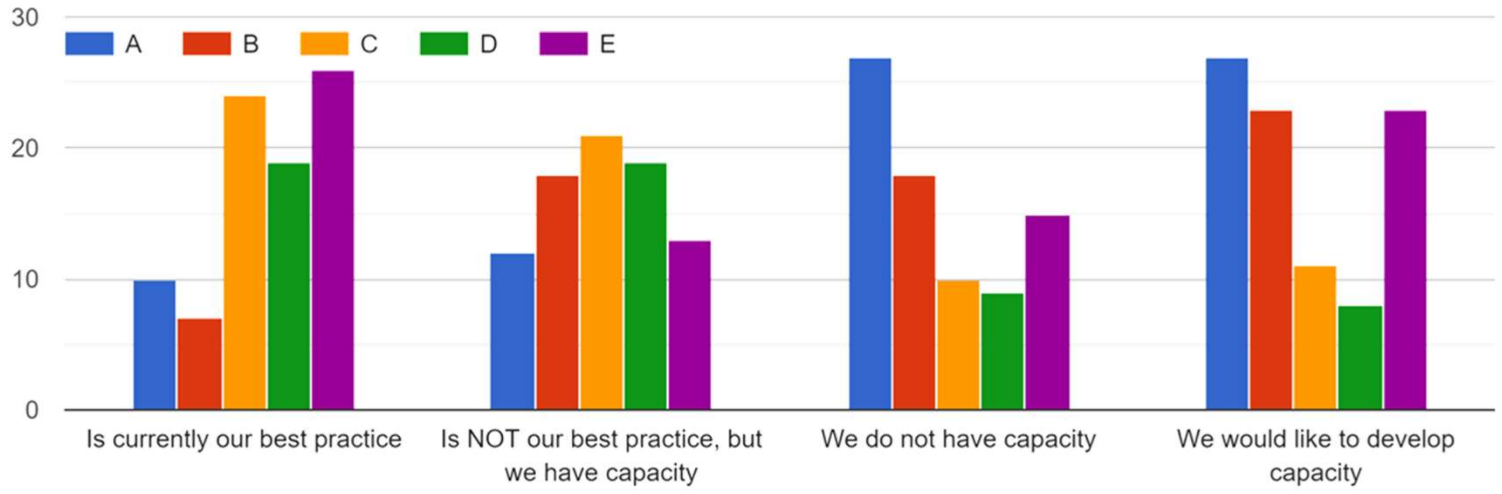
EMBL-EBI

Community Survey

All Atlantic Ocean Microbiome Sampling

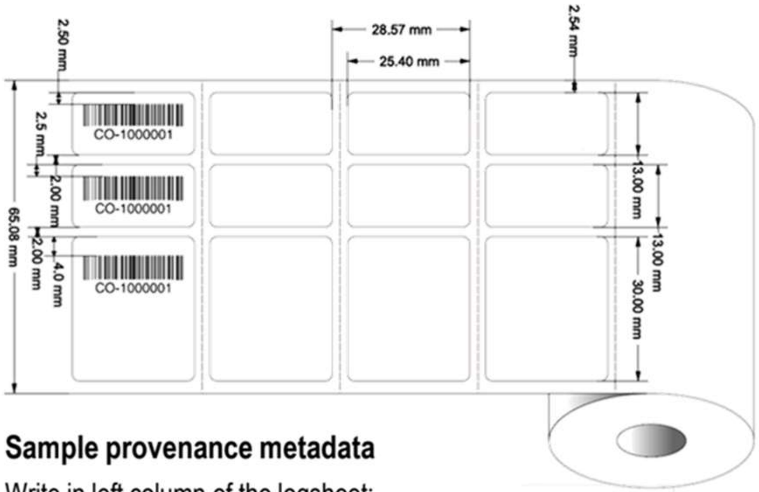
Ocean Sampling Day





EMBL-EBI

Collecting Metadata – unique identifiers



Sample provenance metadata

Write in left column of the logsheet:

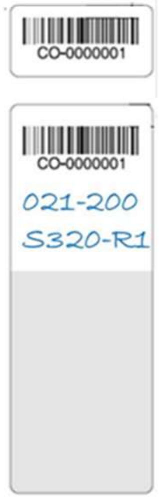
1. Sampling depth (m)
2. Replicate # or Control #

Write on the large barcode sticker:

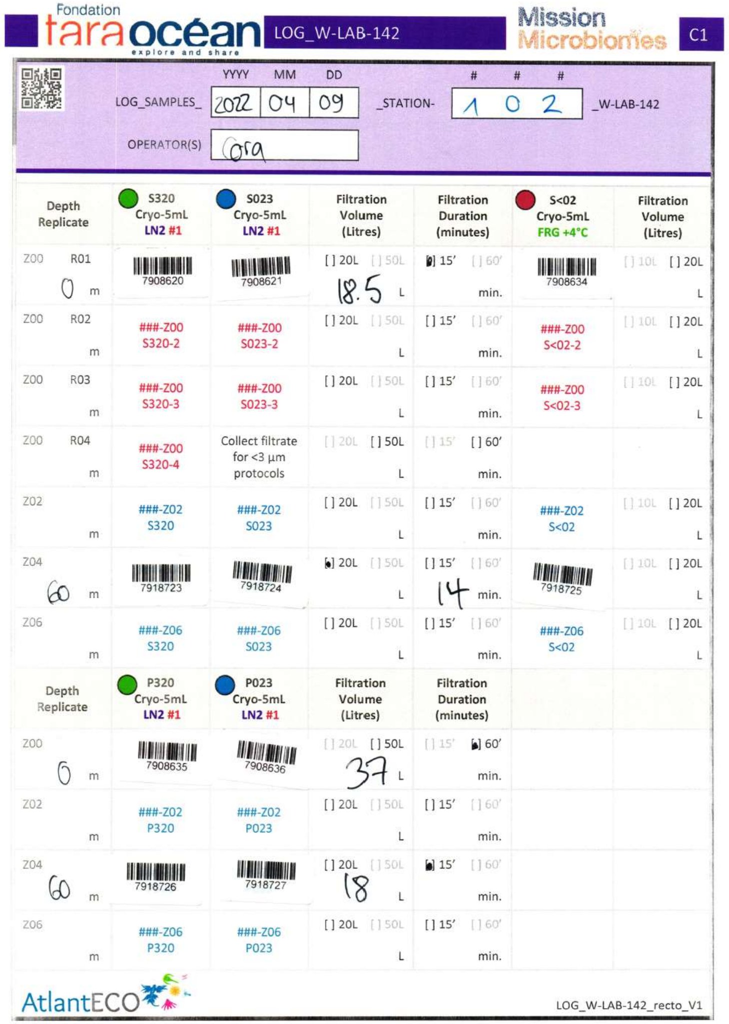
1. Station ID (e.g. 021)
2. Sampling depth (e.g. 200 m)
3. Protocol label (e.g. S320)
4. Replicate # or Control # (e.g R1)

Fix the large barcode sticker on the sample container

Fix the corresponding small barcode sticker on the logsheet, in the appropriate protocol column



Collecting Metadata – provenance & context



Fondation taraocéan explore and share | LOG_W-LAB-142 | Mission Microbiomes | C1

YYYY MM DD	LOG_SAMPLES_ 2022 04 09	_STATION- 1 0 2 _W-LAB-142
OPERATOR(S)	Cora	

Depth Replicate	● S320 Cryo-5mL LN2 #1	● S023 Cryo-5mL LN2 #1	Filtration Volume (Litres)	Filtration Duration (minutes)	● S<02 Cryo-5mL FRG +4°C	Filtration Volume (Litres)
Z00 R01 0 m	7908620	7908621	[ ] 20L [ ] 50L 18.5 L	[●] 15' [ ] 60' min.	7908634	[ ] 10L [ ] 20L L
Z00 R02 m	###-Z00 S320-2	###-Z00 S023-2	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.	###-Z00 S<02-2	[ ] 10L [ ] 20L L
Z00 R03 m	###-Z00 S320-3	###-Z00 S023-3	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.	###-Z00 S<02-3	[ ] 10L [ ] 20L L
Z00 R04 m	###-Z00 S320-4	Collect filtrate for <3 µm protocols	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.		
Z02 m	###-Z02 S320	###-Z02 S023	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.	###-Z02 S<02	[ ] 10L [ ] 20L L
Z04 60 m	7918723	7918724	[●] 20L [ ] 50L L	[ ] 15' [ ] 60' 14 min.	7918725	[ ] 10L [ ] 20L L
Z06 m	###-Z06 S320	###-Z06 S023	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.	###-Z06 S<02	[ ] 10L [ ] 20L L

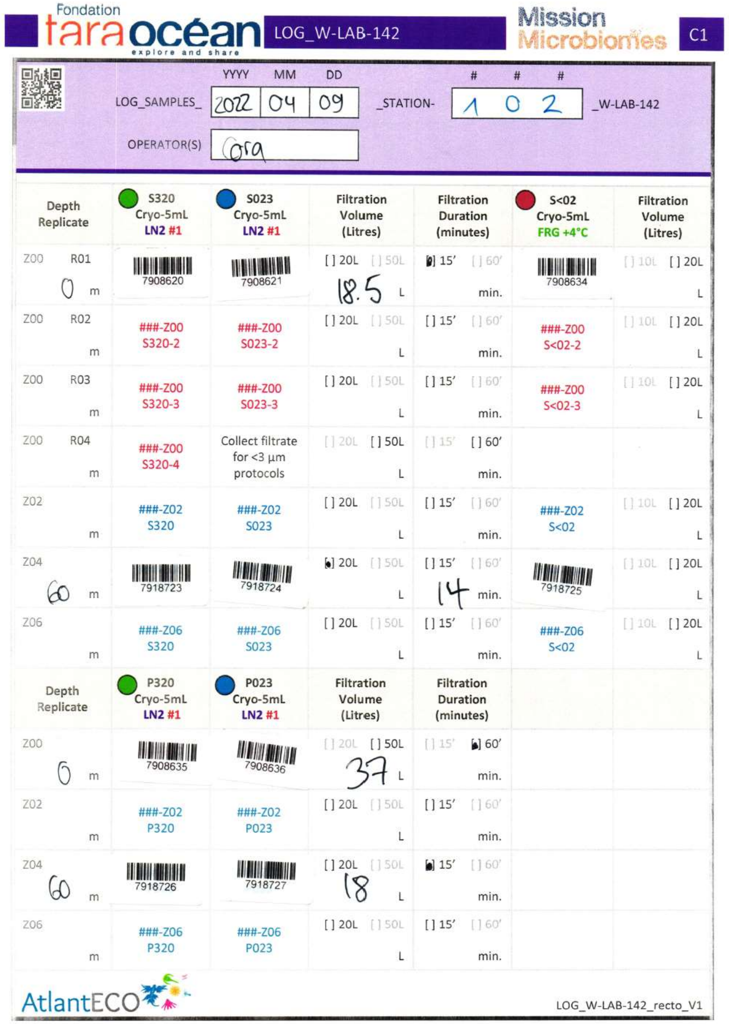
Depth Replicate	● P320 Cryo-5mL LN2 #1	● P023 Cryo-5mL LN2 #1	Filtration Volume (Litres)	Filtration Duration (minutes)
Z00 0 m	7908635	7908636	[ ] 20L [ ] 50L 37 L	[ ] 15' [●] 60' min.
Z02 m	###-Z02 P320	###-Z02 P023	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.
Z04 60 m	7918726	7918727	[ ] 20L [ ] 50L 18 L	[●] 15' [ ] 60' min.
Z06 m	###-Z06 P320	###-Z06 P023	[ ] 20L [ ] 50L L	[ ] 15' [ ] 60' min.

AtlantECO

LOG_W-LAB-142_recto_V1

EMBL-EBI

Collecting Metadata – structured checklists



Fondation tara océan explore and share

LOG_W-LAB-142

Mission Microbiomes

C1

LOG_SAMPLES_ YYYY 2022 MM 04 DD 09 _STATION- # 1 # 0 # 2 _W-LAB-142

OPERATOR(S) Cora

Depth Replicate	● S320 Cryo-5mL LN2 #1	● S023 Cryo-5mL LN2 #1	Filtration Volume (Litres)	Filtration Duration (minutes)	● S<02 Cryo-5mL FRG +4°C	Filtration Volume (Litres)
Z00 R01 0 m	7908620	7908621	[ ] 20L [ ] 50L 18.5 L	[●] 15’ [ ] 60’ min.	7908634	[ ] 10L [ ] 20L L
Z00 R02 m	###-Z00 S320-2	###-Z00 S023-2	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.	###-Z00 S<02-2	[ ] 10L [ ] 20L L
Z00 R03 m	###-Z00 S320-3	###-Z00 S023-3	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.	###-Z00 S<02-3	[ ] 10L [ ] 20L L
Z00 R04 m	###-Z00 S320-4	Collect filtrate for <3 µm protocols	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.		
Z02 m	###-Z02 S320	###-Z02 S023	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.	###-Z02 S<02	[ ] 10L [ ] 20L L
Z04 60 m	7918723	7918724	[●] 20L [ ] 50L L	[ ] 15’ [ ] 60’ 14 min.	7918725	[ ] 10L [ ] 20L L
Z06 m	###-Z06 S320	###-Z06 S023	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.	###-Z06 S<02	[ ] 10L [ ] 20L L
Depth Replicate	● P320 Cryo-5mL LN2 #1	● P023 Cryo-5mL LN2 #1	Filtration Volume (Litres)	Filtration Duration (minutes)		
Z00 0 m	7908635	7908636	[ ] 20L [ ] 50L 37 L	[ ] 15’ [●] 60’ min.		
Z02 m	###-Z02 P320	###-Z02 P023	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.		
Z04 60 m	7918726	7918727	[ ] 20L [ ] 50L 18 L	[●] 15’ [ ] 60’ min.		
Z06 m	###-Z06 P320	###-Z06 P023	[ ] 20L [ ] 50L L	[ ] 15’ [ ] 60’ min.		

AtlantECO

LOG_W-LAB-142_recto_V1

attribute	Format / units	comment
sample id	SAMEA0000000	
sample label	Alpha-numeric, e.g. “project_date-time_station_environment_size-fraction_method”	human readable and meaningful label
sampling design, label(s)	alpha-numeric	campaign, station, site, transect, etc.
sampling device	alpha-numeric	device name & specifications
operator	alpha-numeric	initials or full name
sampling date and time	yyyy-mm-dd T hh:mm	in UTC
latitude	N/S dd.dddddd or N/S dd mm.mmm	ultimately in decimal degree N
longitude	E/W ddd.dddddd or E/S ddd mm.mmm	ultimately in decimal degree E
elevation, depth below soil surface	cm	ultimately in metre
elevation, depth below sediment surface	cm	ultimately in metre
elevation, depth below water surface	m	
elevation, altitude above sea level	m	
methodological details	alpha-numeric	e.g. processing time & volume

EMBL-EBI

Data sharing best practices

Permanent archives selected for the different data types:

- BioSamples for metadata (https://www.ebi.ac.uk/biosamples/)
- ENA for genomics data (http://www.ebi.ac.uk/ena)
- MGnify for metagenomic data (https://www.ebi.ac.uk/metagenomics/)
- PRIDE for proteomics data (https://www.ebi.ac.uk/pride/)
- Metabolights for metabolomics data (https://www.ebi.ac.uk/metabolights/)
- BioImage Archive for imaging data (https://www.ebi.ac.uk/bioimage-archive/)

Data workflow

primary data
curation & archiving

annotated data
analysis & archiving

Environmental
EMODnet, Copernicus, etc.

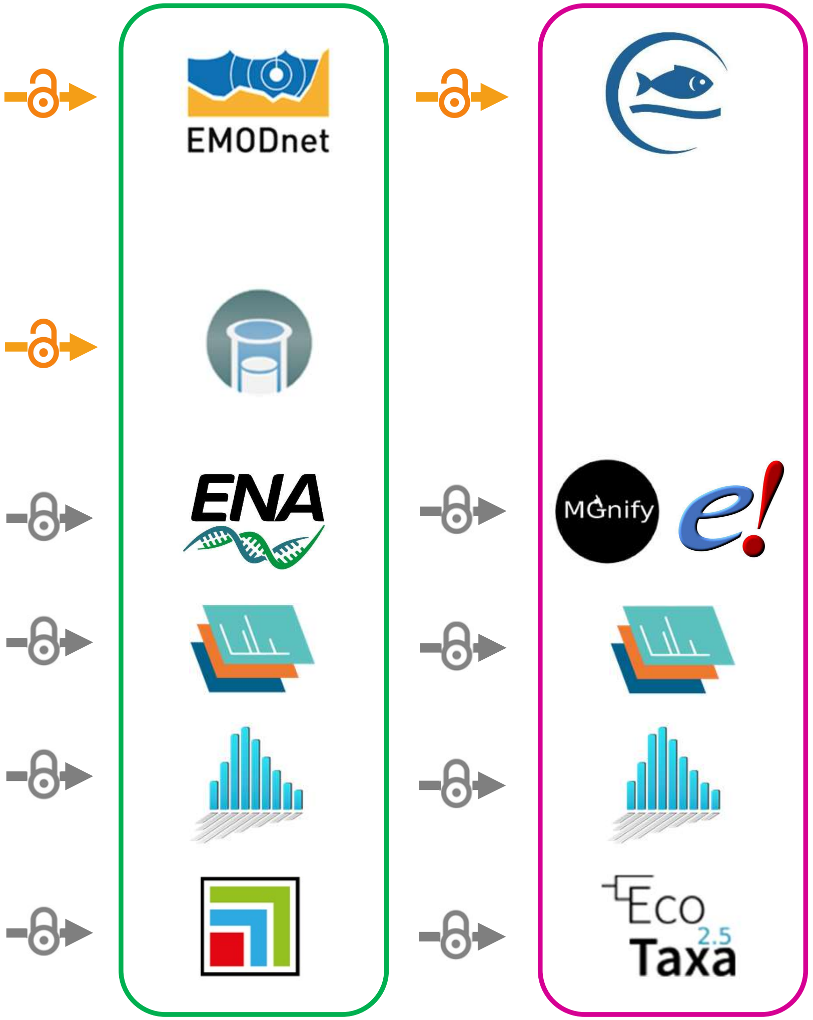
Provenance & Envir. Context
BioSamples

Genomics
ENA, Mgnify & Ensembl

Proteomics
PRIDE

Metabolomics
Metabolights

Imaging
BioImage archive & EcoTaxa



EMBL-EBI

Data workflow

Environmental
EMODnet, Copernicus, etc.

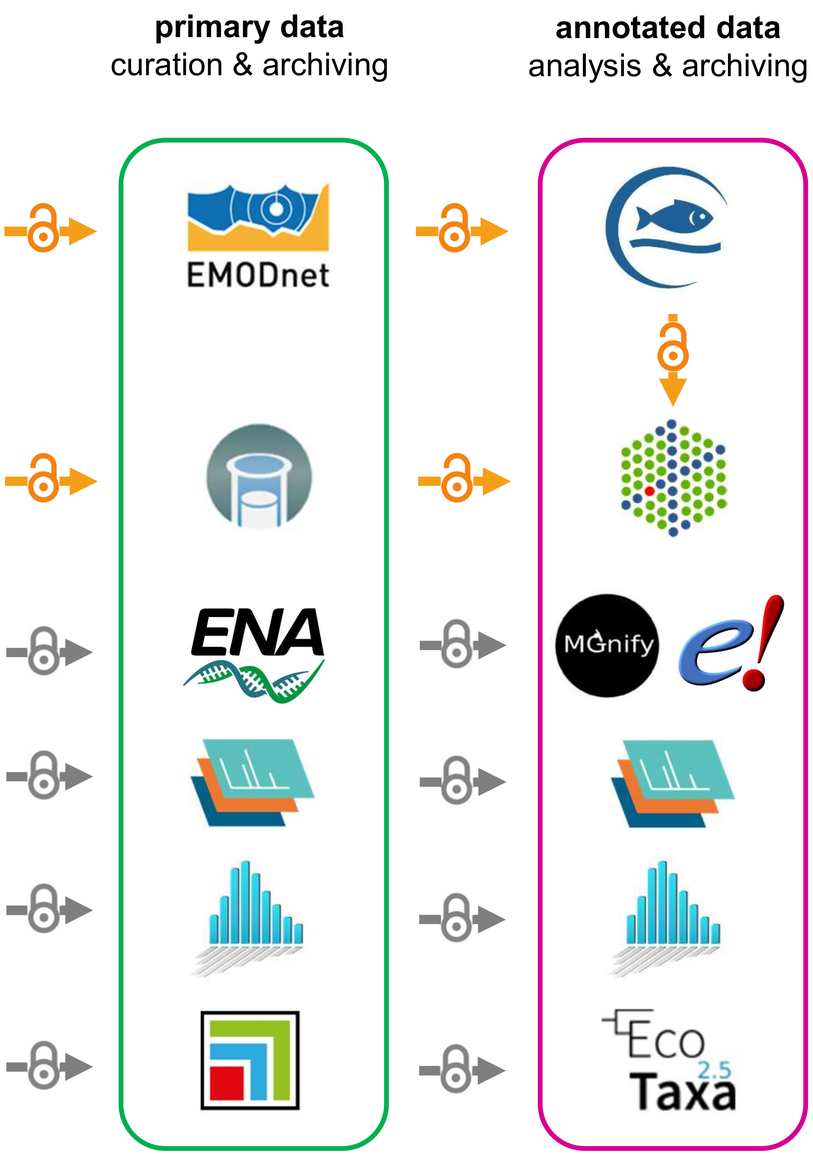
Provenance & Envir. Context
BioSamples

Genomics
ENA, Mgnify & Ensembl

Proteomics
PRIDE

Metabolomics
Metabolights

Imaging
BioImage archive & EcoTaxa



primary data
curation & archiving

annotated data
analysis & archiving

Harmonised context

- Georeferences
- Taxonomy
- Marine domain

- Longhurst Biogeo. Provinces
- IHO Sea Areas
- Oceanographic Products
- Astronomical Products

- Countries & Territories
- Exclusive Economic Zones
- Nagoya requirements
- ABS regulations

EMBL-EBI

Data workflow

Environmental
EMODnet, Copernicus, etc.

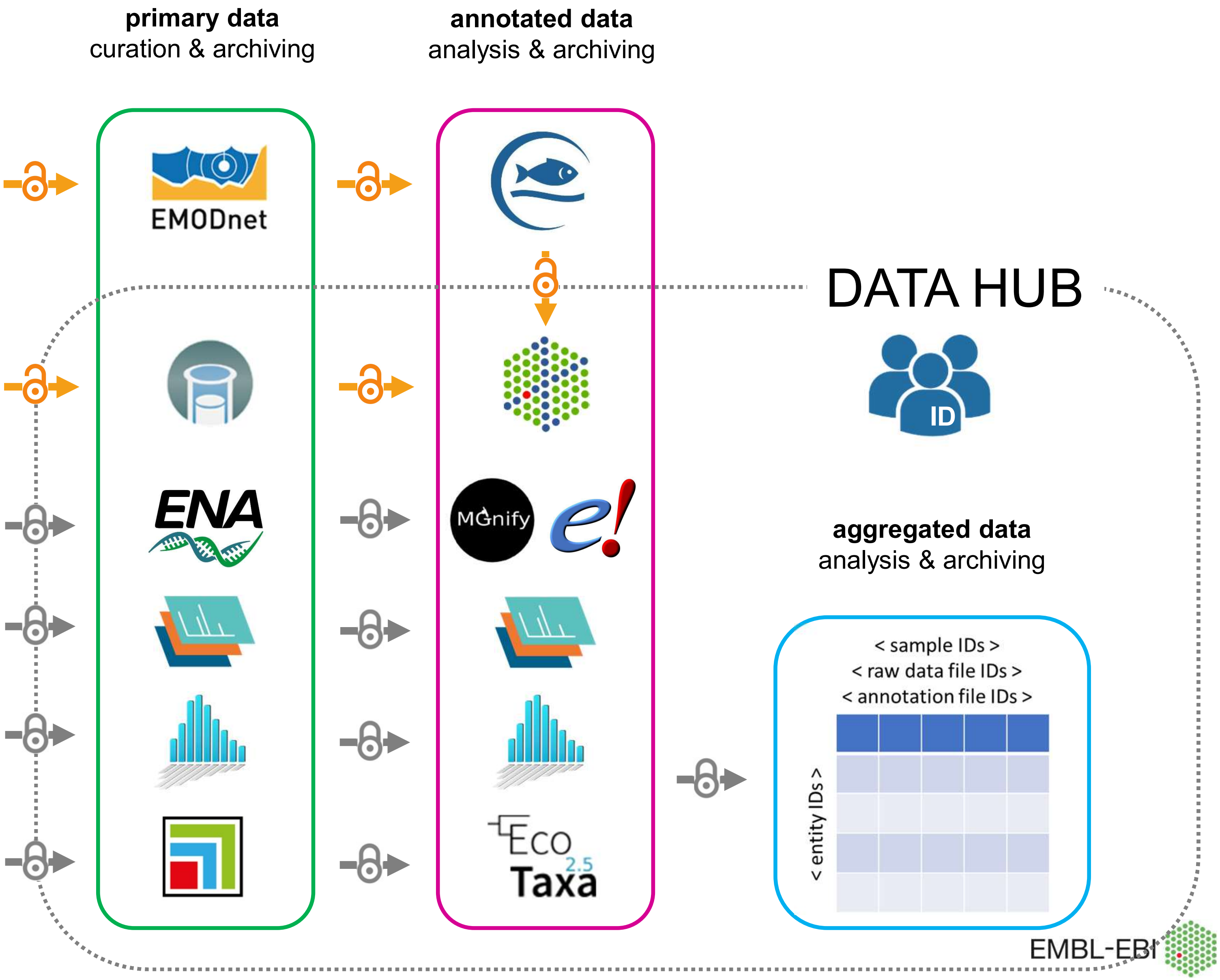
Provenance & Envir. Context
BioSamples

Genomics
ENA, Mgnify & Ensembl

Proteomics
PRIDE

Metabolomics
Metabolights

Imaging
BioImage archive & EcoTaxa



primary data
curation & archiving

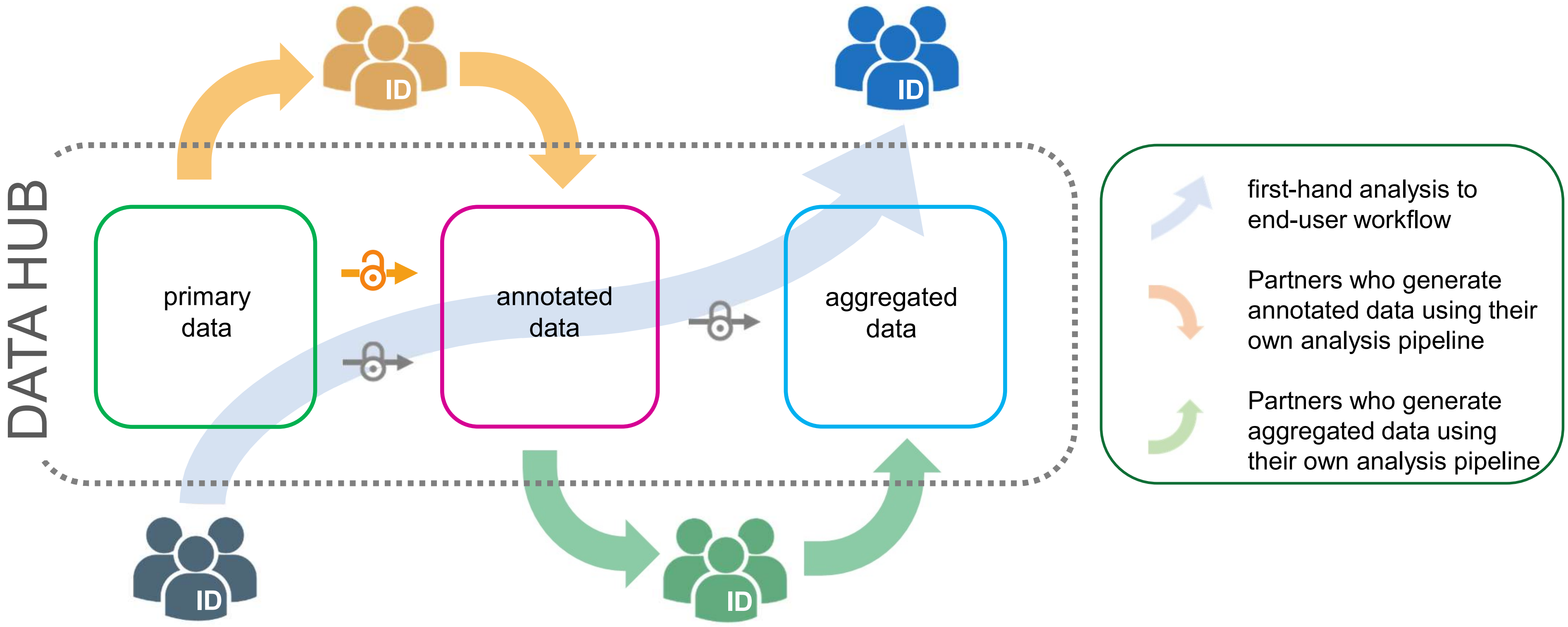
annotated data
analysis & archiving

DATA HUB

aggregated data
analysis & archiving

< sample IDs >
< raw data file IDs >
< annotation file IDs >
< entity IDs >

Data Hub workflow



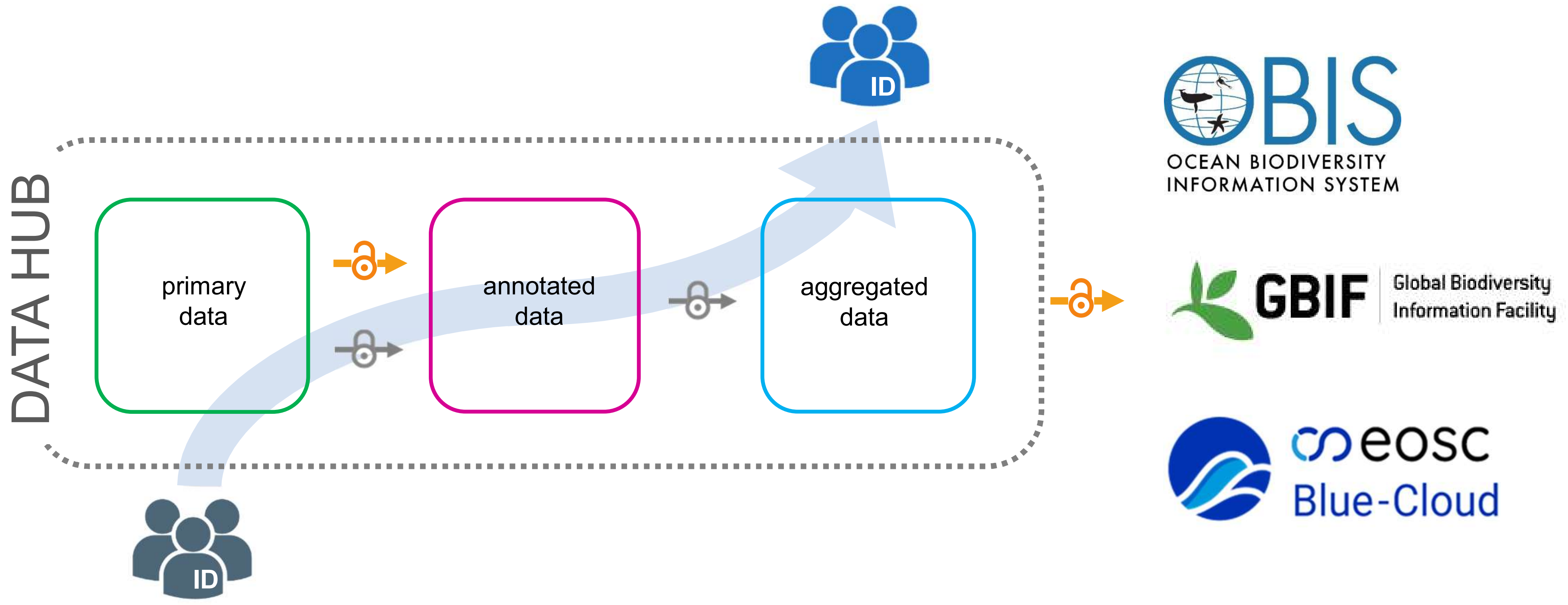
first-hand analysis to end-user workflow

Partners who generate annotated data using their own analysis pipeline

Partners who generate aggregated data using their own analysis pipeline

EMBL-EBI

Data Hub workflow

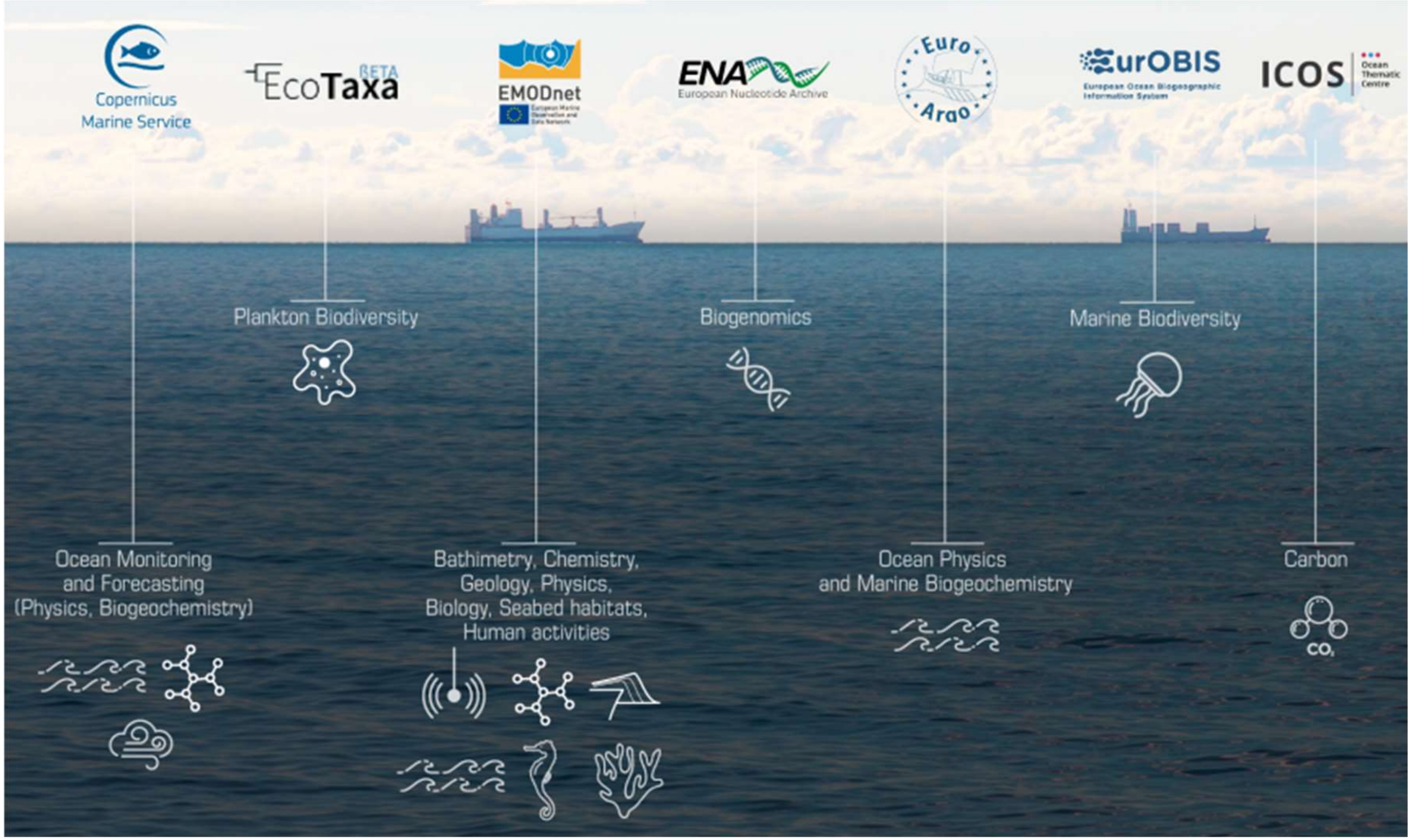

DATA HUB
primary data
annotated data
aggregated data
ID
ID
OBIS
OCEAN BIODIVERSITY INFORMATION SYSTEM
GBIF Global Biodiversity Information Facility
eosc Blue-Cloud

EMBL-EBI

eosc | Blue-Cloud

https://blue-cloud.org/

Discovery & Access

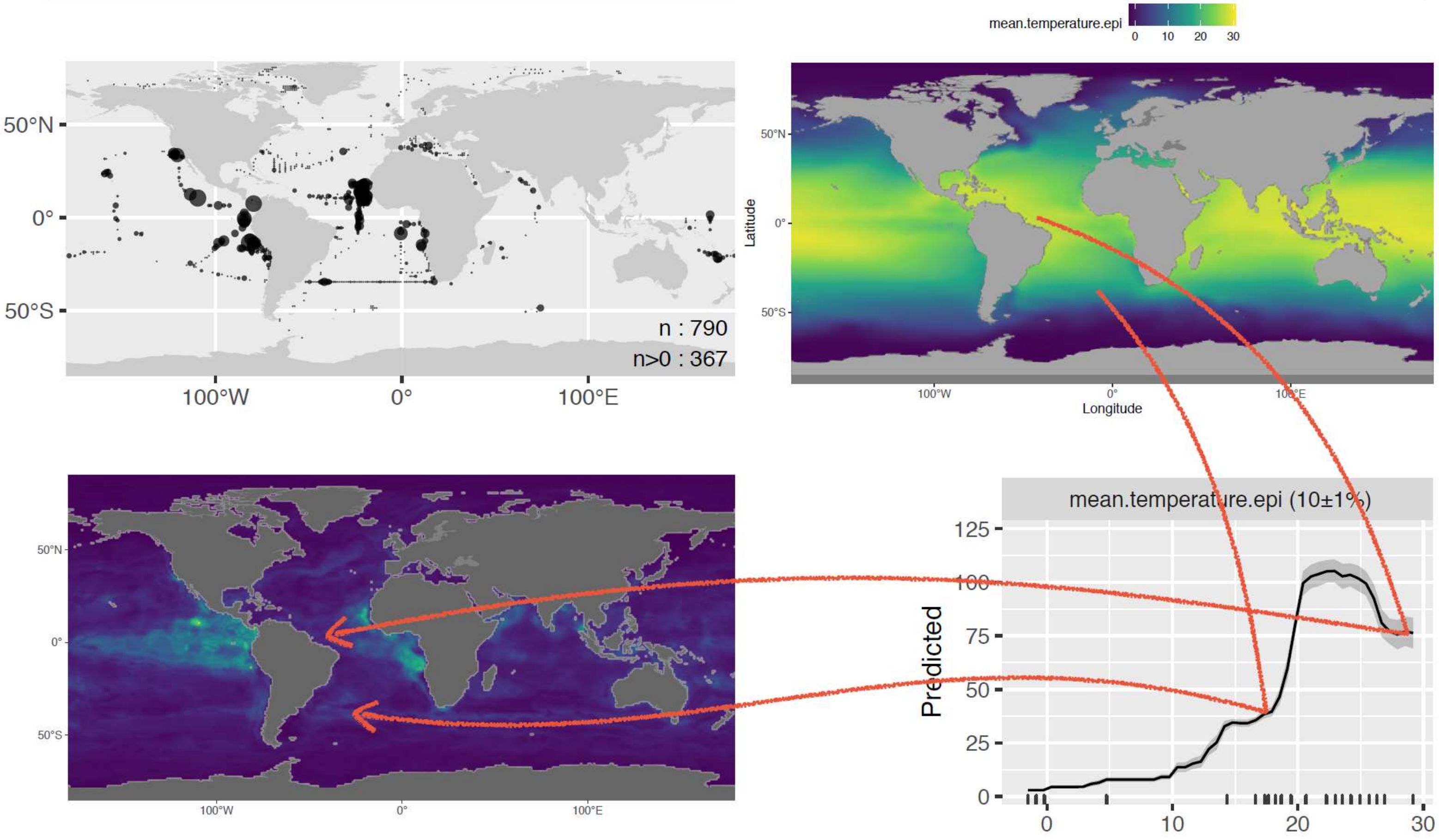

Copernicus Marine Service
EcoTaxa BETA
EMODnet
ENA European Nucleotide Archive
Euro Argo
EurOBIS European Ocean Biogeographic Information System
ICOS Ocean Thematic Centre
Plankton Biodiversity
Biogenomics
Marine Biodiversity
Ocean Monitoring and Forecasting (Physics, Biogeochemistry)
Bathimetry, Chemistry, Geology, Physics, Biology, Seabed habitats, Human activities
Ocean Physics and Marine Biogeochemistry
Carbon
CO₂

eosc | Blue-Cloud

https://blue-cloud.org/

Discovery & Access

Virtual Labs & Workbenches



EMBL-EBI

eosc | Blue-Cloud

https://blue-cloud.org/

Discovery & Access

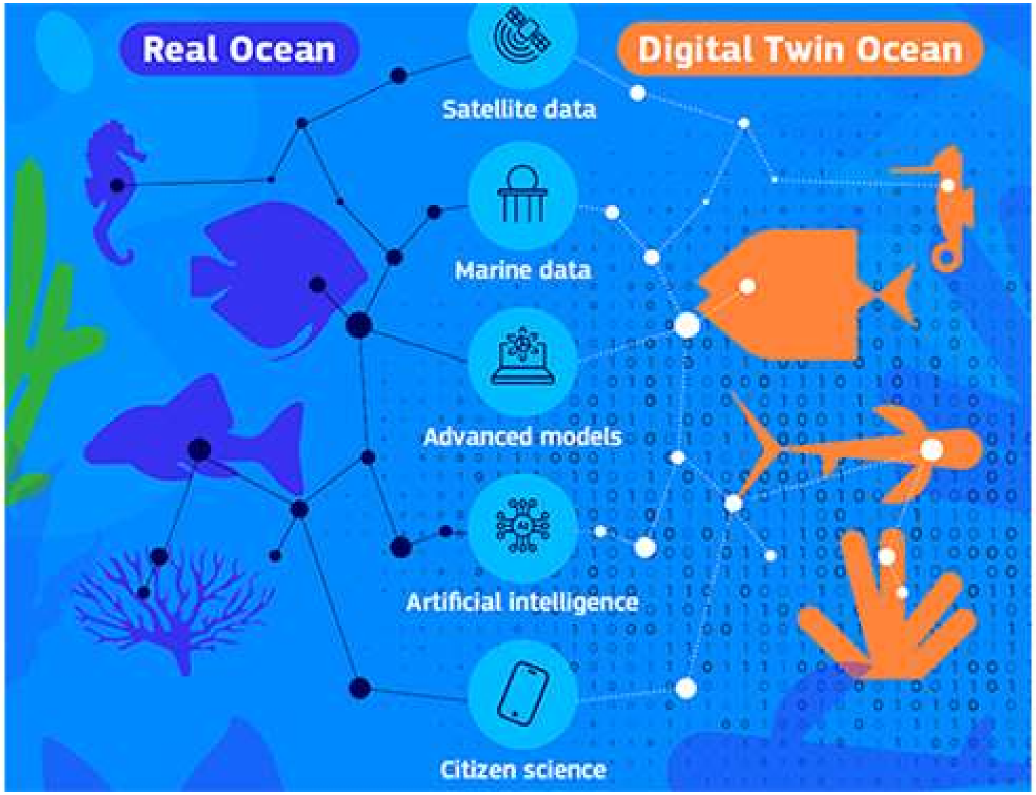
Virtual Environment

Publishing Services

EMODnet

zenodo

OpenAIRE



EMBL-EBI

Publication best practices

Publication
Methods
Material collection
Samples
Software
Dataset
Presentation
Poster

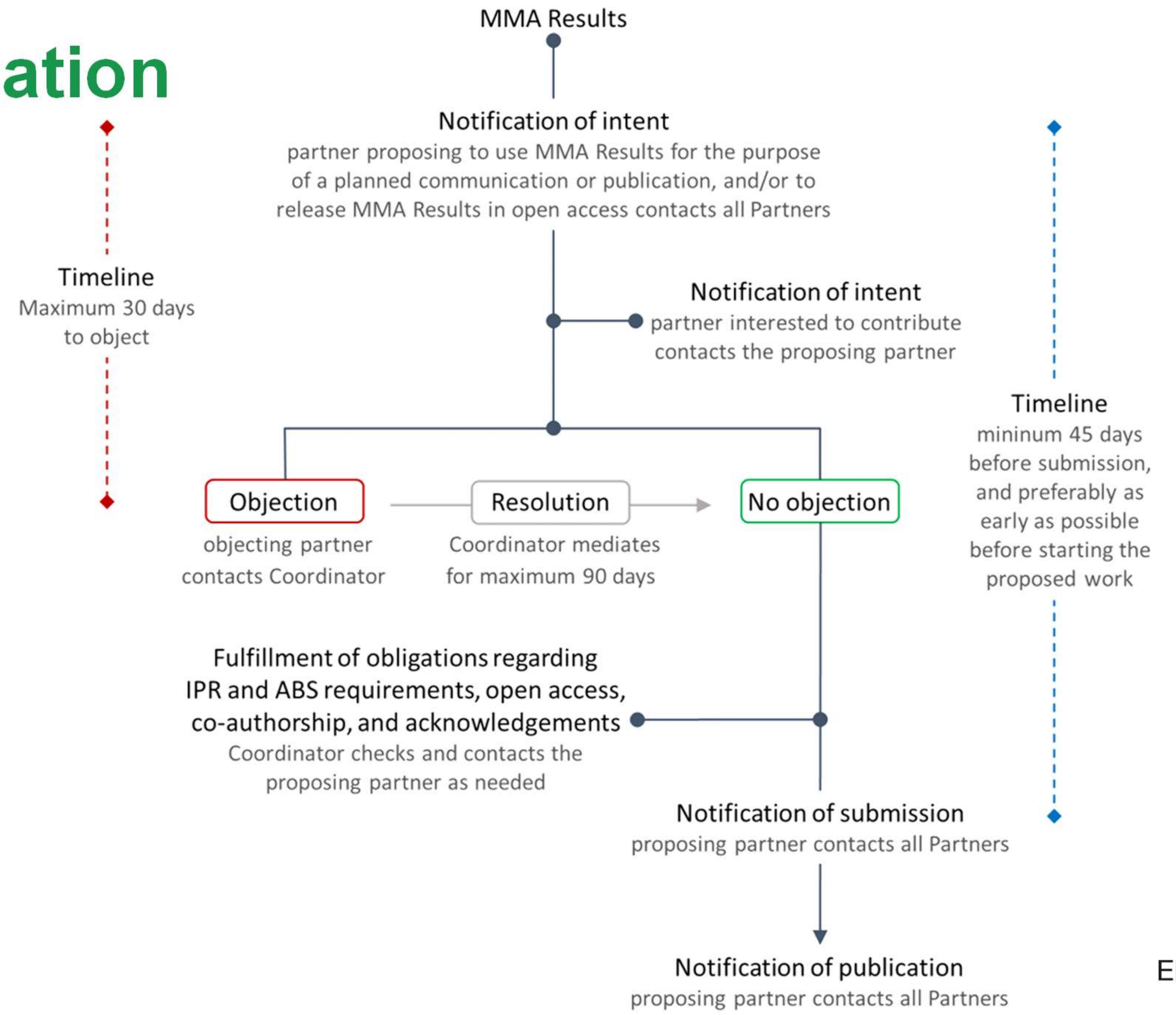
Exemplar initiatives

- Tara Oceans
- Tara Pacific
- AtlantECO

Best Practices

- Jointly owned results
- Early notification of intent to use “jointly owned results”
- Co-authorship
- Open Access publication

Early notification



MMA Results

Notification of intent

partner proposing to use MMA Results for the purpose of a planned communication or publication, and/or to release MMA Results in open access contacts all Partners

Timeline

Maximum 30 days to object

Notification of intent

partner interested to contribute contacts the proposing partner

Timeline

mininum 45 days before submission, and preferably as early as possible before starting the proposed work

Objection

objecting partner contacts Coordinator

Resolution

Coordinator mediates for maximum 90 days

No objection

Fulfillment of obligations regarding IPR and ABS requirements, open access, co-authorship, and acknowledgements

Coordinator checks and contacts the proposing partner as needed

Notification of submission

proposing partner contacts all Partners

Notification of publication

proposing partner contacts all Partners

EMBL-EBI

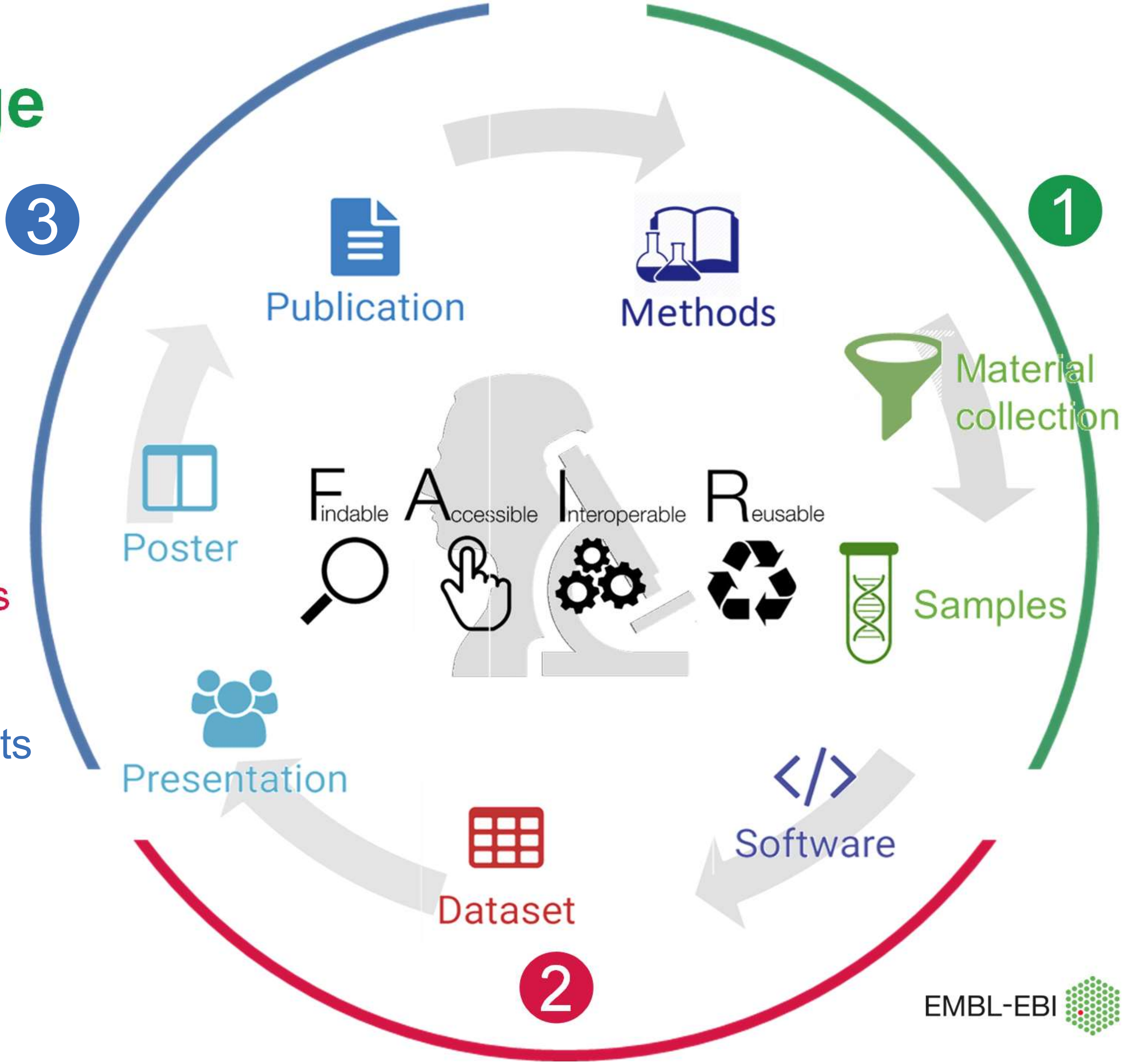
Take home message

Use standard methods
ADAPT
Collect structured metadata

Deposit metadata & data
EARLY
Progressive closed > open access

Communicate your research intents
EARLY
Seek collaboration

Publish in open access



EMBL-EBI

Thank you for your kind attention

Stéphane PESANT

Senior marine biocurator

pesant@ebi.ac.uk

EMBL-EBI