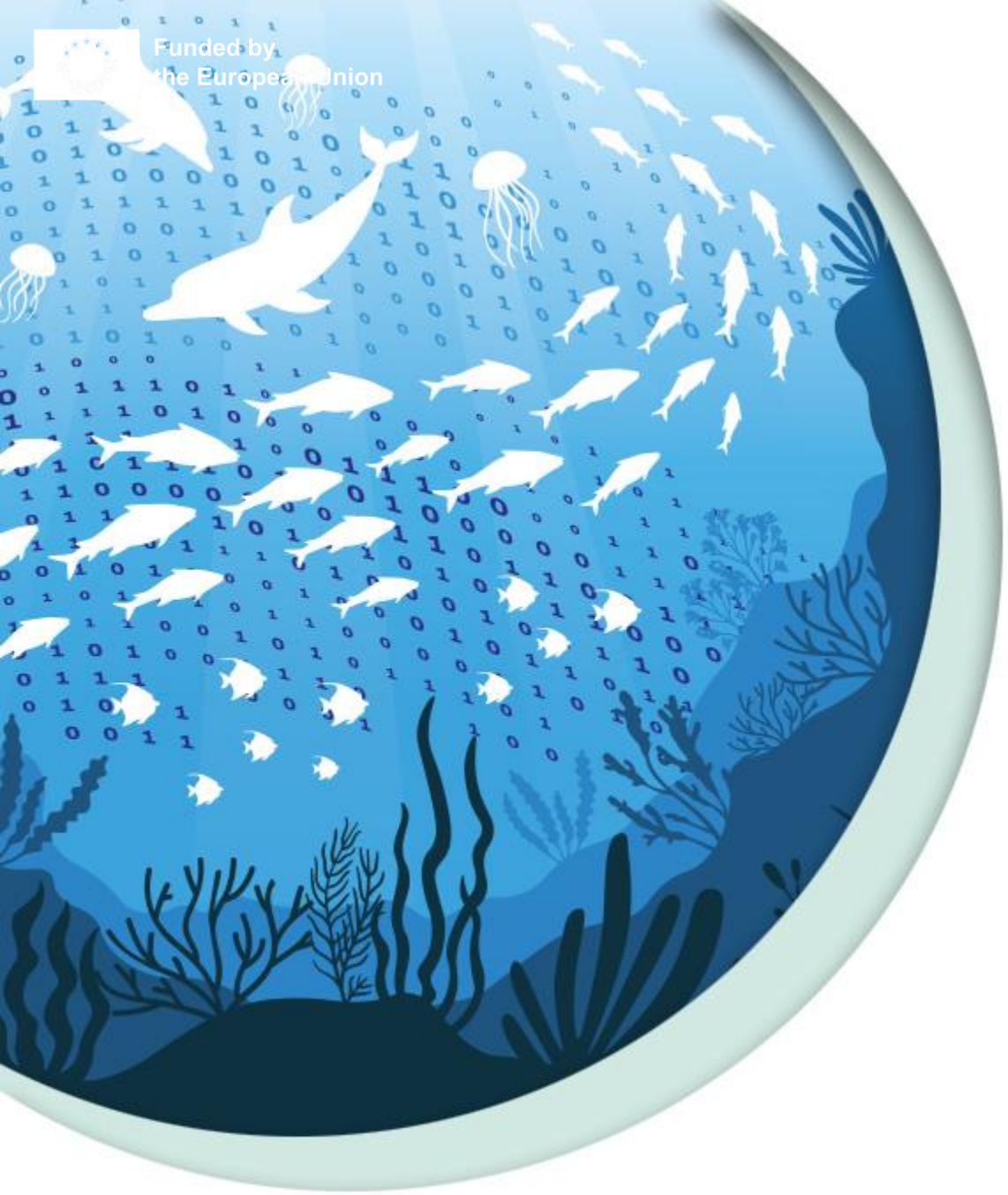
Funded by
the European Union



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

DTO-BioFlow data training workshop:

FAIR data: why and how

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Research data management

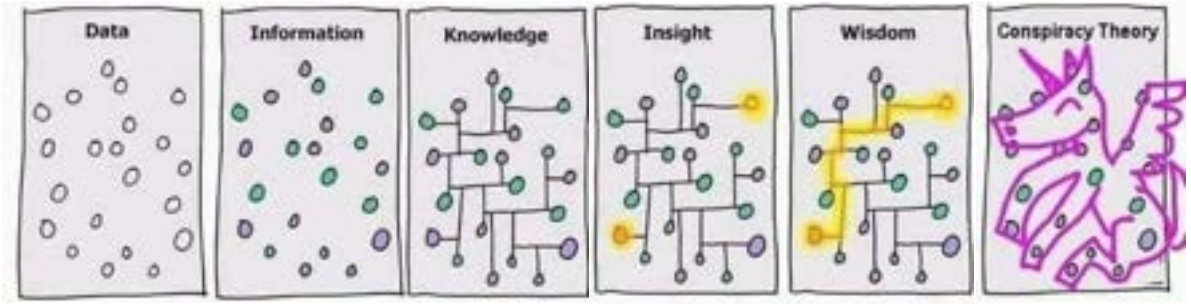
Why it is important

FAIR & Open data

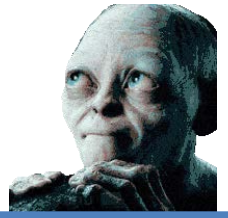
Research data cycle

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean



Why it is important



"Data is a **precious** thing and will last longer than the systems themselves."

Tim Berners-Lee, inventor of the World Wide Web

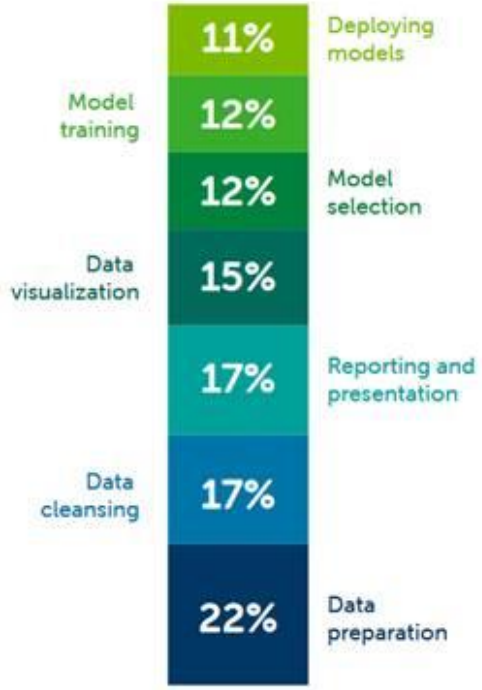
VLIZ

Order, chaos or organised chaos?

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

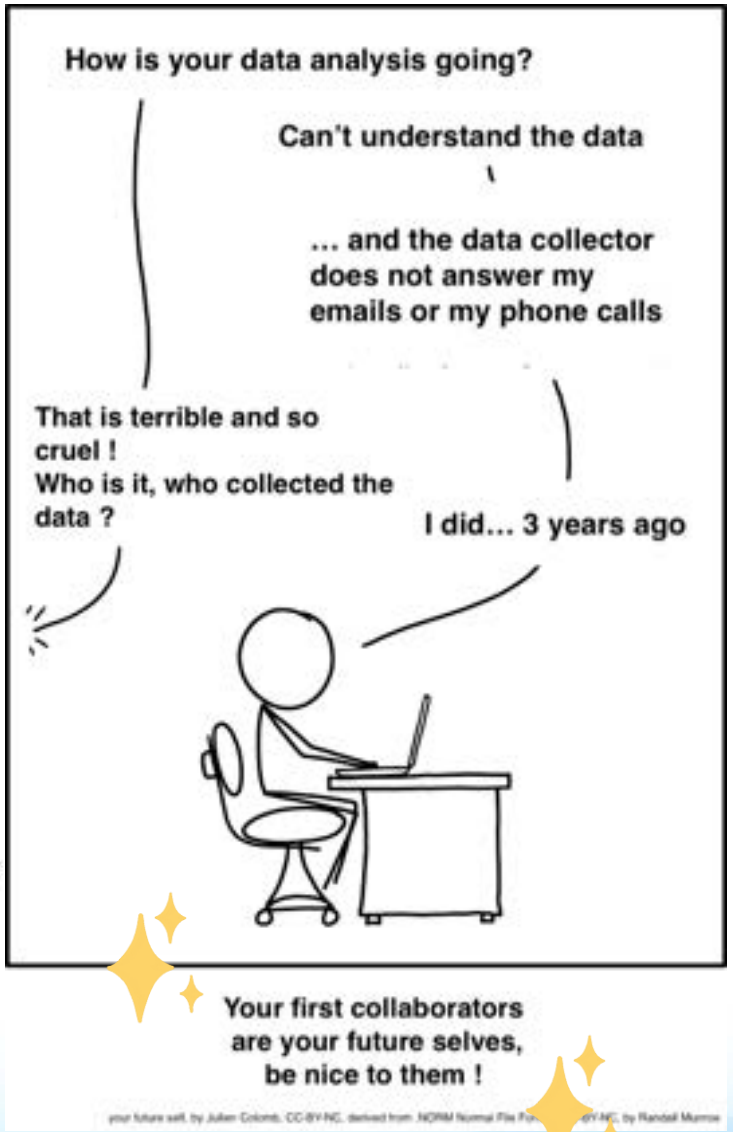
"We are all data providers and data users"

under the current system. Students in PhD programmes spend up to 80% of their time on 'data munging', fixing formatting and minor mistakes to make data suitable for analysis – wasting time and talent. With 400 such students, that would amount to a monetary waste equivalent to the salaries of 200 full-time employees, at minimum. So, hiring 20 professional data stewards to cut time lost to data wrangling would boost effective research capacity. Many top



n = 2,030

We asked our respondents how much time they spend on each of the above tasks, and for each item, enter a number representing the percentage of time spent on each task relative to the other tasks on this list. The percentage values had to add up to 100.



More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from *Nature*'s survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604).https://doi.org/10.1038/533452a
Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. Nature, 578(7796), 491-491. https://doi.org/10.1038/d41586-020-00505-7
Anaconda (2021). 2021 State of Data Science report. https://www.anaconda.com/resources/whitepapers/state-of-data-science-2021

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Data sharing: benefits of sharing data

Personal benefits

More references & credits to your work

Career recognition

Collaborations

Moral obligations

Efficient use of public resources

Data are unique & come with a high cost

Facilitates data finding & re-use

→ New research & new insights

Better data leads to better research

→ Improved decisions-making

→ Increased transparency & trust in science

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Research data management

Why it is important

FAIR & Open data

Research data cycle

VLIZ

DTO-BioFl
Integration of biodiversity monitoring data into the Digital Twin Oc

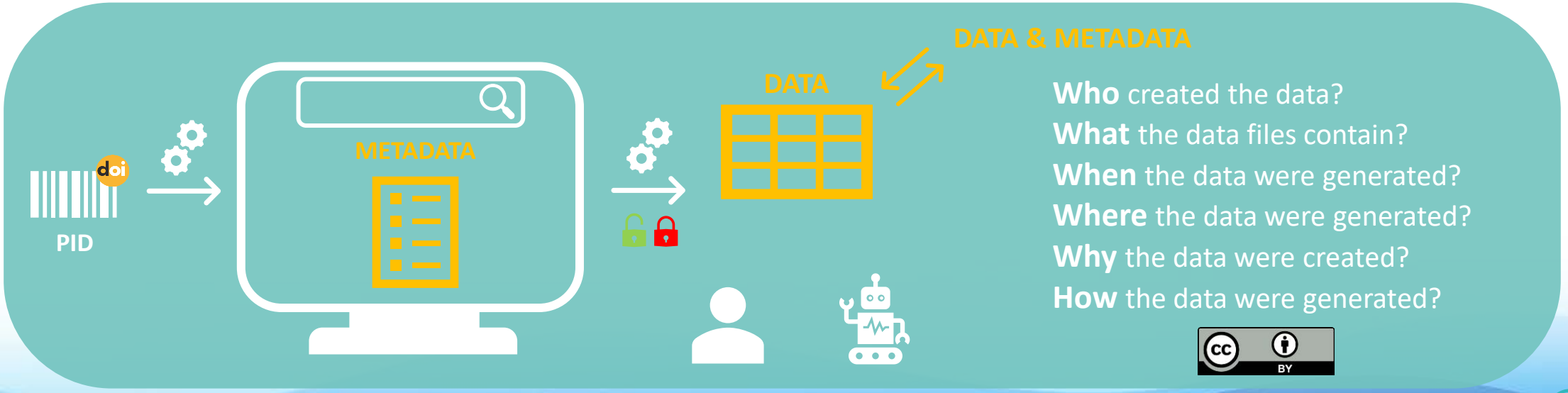
F indable A ccessible I nteroperable R eusable

- Rich metadata & available online
- Persistent identifier

- Retrievable
- Accessible ≠ OPEN
- Authentication & authorisation steps
- Metadata should always be accessible

- Machine readable components
- Open formats
- Recognized standards
- Linked data
- Integration ready

- Data ‘provenance’
- Data usage licence



Ghent University Data Stewards (2020). Knowledge clip: FAIR data principles. Available at https://youtu.be/2uZxFu9SFi8

VLIZ

FAIR data

It is a spectrum

≠ Open data

Open data is data that anyone can access, use & share

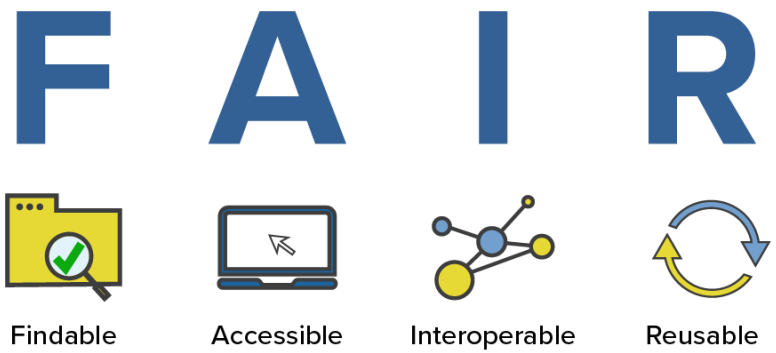
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Responsibilties

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Responsibilties



VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

The circle of life
Research data management

Why it is important

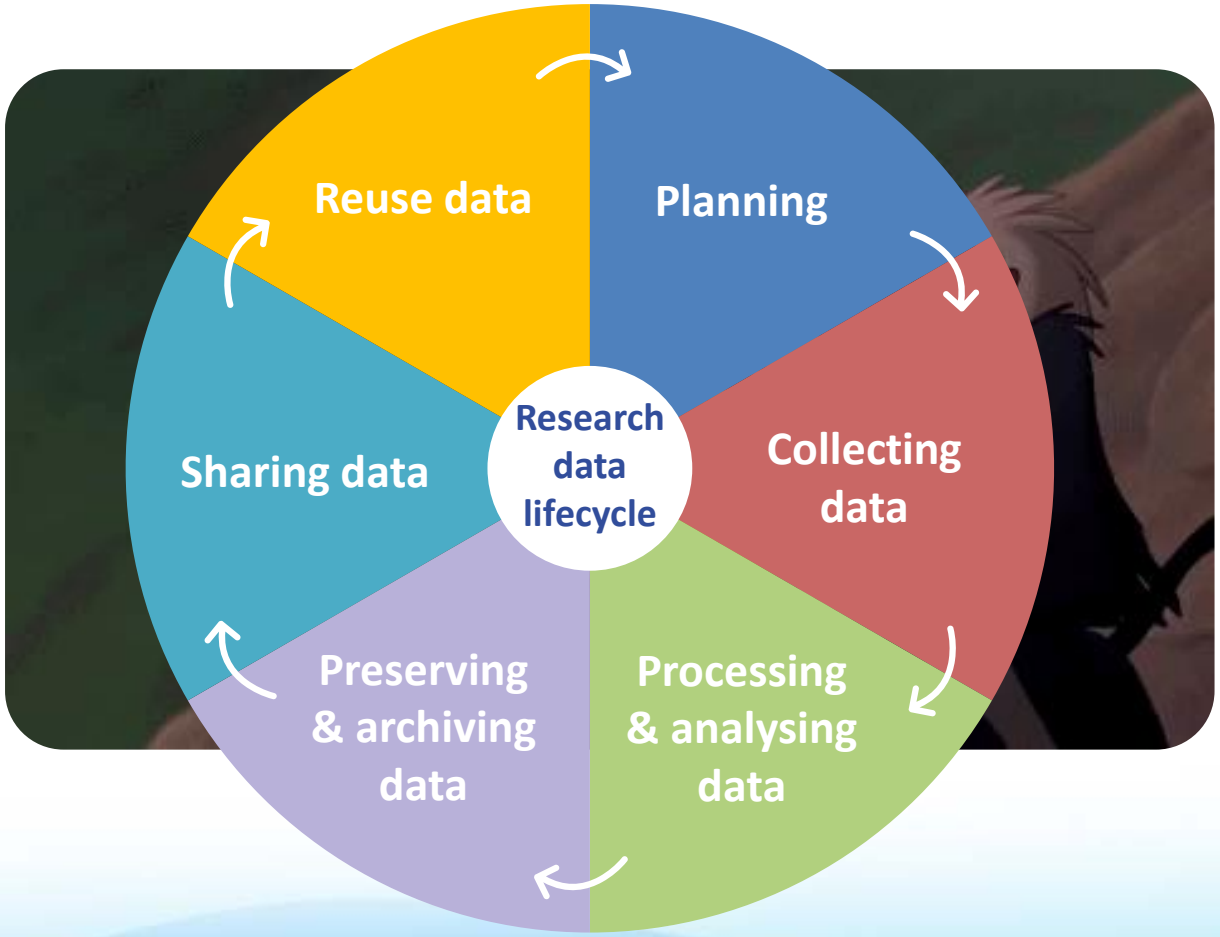
FAIR & Open data

Research data cycle

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

The circle of ~~life~~ data



VLIZ

RDM in practice!

Planning - DMP

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Planning

Data Management Plan

What?

- How data will be handled **during & after** a research project
- Formal & “living” document

Why?

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Planning

Data Management Plan

Content of DMP

DMP templates

DMPonline.be

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting data

- Metadata
- Controlled vocabularies
- Standards

VLIZ

Metadata and Documentation

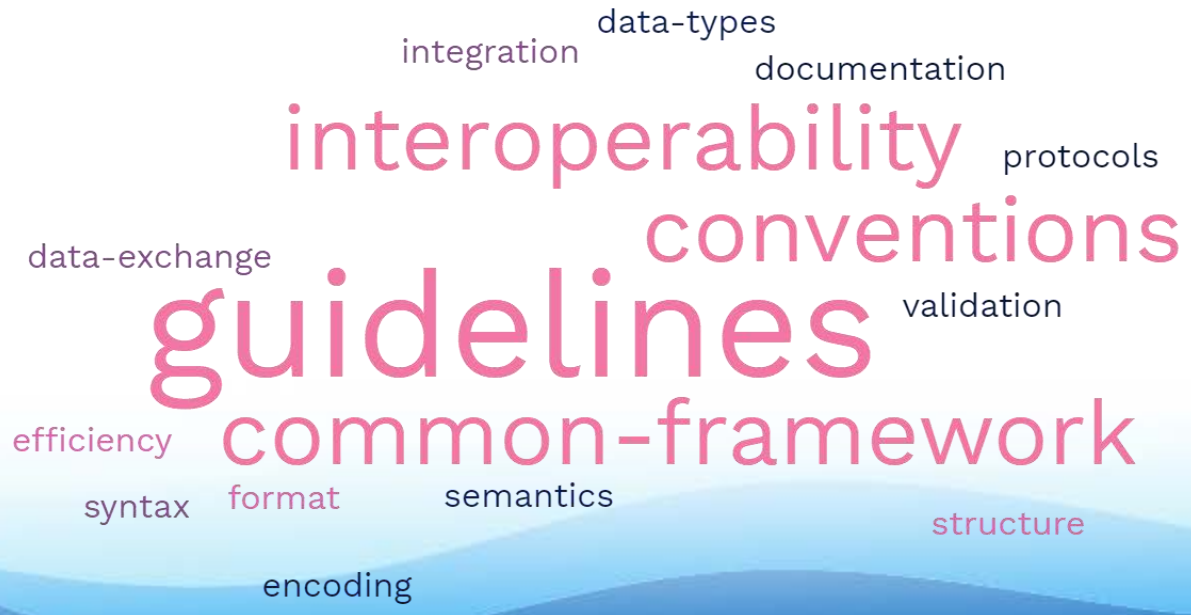
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting

(Meta)Data standards

Global & multidisciplinary standards:

"Set of guidelines or rules that specify how data should be structured, formatted, and represented to ensure consistency, interoperability, and efficient data exchange"



VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting

Data standards

Global & multidisciplinary standards:

DwC-A and DwC

= Darwin Core

EML

= Ecological Metadata Language

eventID	parentEventID	eventDate	decimalLongitude	decimalLatitude
site_1			54.7943	16.9425
zone_1	site_1			
zone_2	site_1			
zone_3	site_1			
quadrat_1	zone_1	2019-01-02		
transect_1	zone_2	2019-01-03		
transect_2	zone_3	2019-01-04		

id	occurrenceID	scientificName
quadrat_1	occ_1	Ulva rigida
quadrat_1	occ_2	Ulva lactuca
transect_1	occ_3	Plantae
transect_1	occ_4	Plantae
transect_2	occ_5	Gracilaria
transect_2	occ_6	Laurencia

- Basic Metadata
- Geographic Coverage
- Taxonomic Coverage
- Temporal Coverage
- Keywords
- Associated Parties
- Project Data
- Sampling Methods
- Citations
- Collection Data
- External links
- Additional Metadata

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting

Controlled vocabularies

- List of terms where each term means just one thing
- Ensure standardisation

National Oceanography Centre
British Oceanographic Data Centre BODC

- Biomass

Identifier ↑	Preferred label ↑	Alternative label ↑	Definition ↑
SDBIOL09	Dry weight biomass of biological entity specified elsewhere per unit volume of the water body	WaterDryWtBiom_BE007117	The mass measured after drying at elevated temperatures until a stable mass is reached, of an identified biological object described elsewhere in the metadata occurring in a given volume of any body of salt or fresh water.
SDBIOL07	Ash-free dry weight biomass of biological entity specified elsewhere per unit volume of the water body	WaterAshFreeBiom_BE007117	The mass lost on ignition of an identified biological object described elsewhere in the metadata occurring in a given volume of any body of salt or fresh water.
SDBIOL04	Wet weight biomass of biological entity specified elsewhere per unit volume of the water body	WaterWetWtBiom	The mass as caught of an identified biological object described elsewhere in the metadata occurring in a given volume of any body of salt or fresh water.
SDBIOL12	Biomass as carbon of biological entity specified elsewhere per unit volume of the water body by computation	WaterCarbonBiomassConv	The carbon biomass, calculated from the cell counts using literature conversion factors, of an unspecified biological entity in a given volume of any body of salt or fresh water.

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting

Taxonomic standard

WoRMS provides the most authoritative list of names of all marine species globally, ever published

WoRMS
World Register of Marine Species

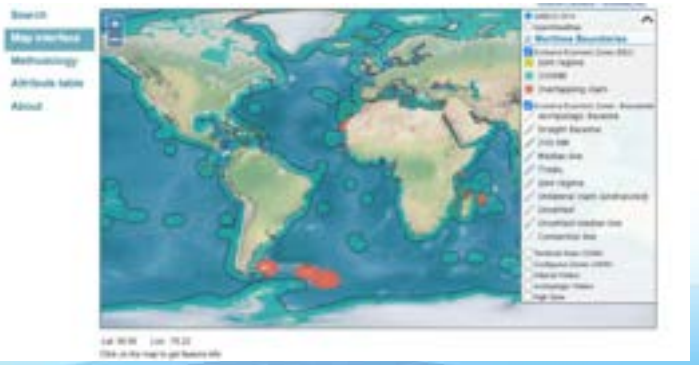
WoRMS taxon details

Glaucus atlanticus Forster, 1777

AphiaID	140022 (urn:lsid:marinespecies.org:taxname:140022)
Classification	Biota > Animalia (Kingdom) > Mollusca (Phylum) > Gastropoda (Class) > Heterobranchia (Subclass) > Euthyneura (Infraclass) > Ringipleura (Subterclass) > Nudipleura (Superorder) > Nudibranchia (Order) > Cladobranchia (Suborder) > Aeolidioidea (Superfamily) > Glaucidae (Family) > *Glaucus* (Genus) > *Glaucus atlanticus* (Species)
Status	accepted

Geographic standard

Standard list of marine georeferenced place names & areas



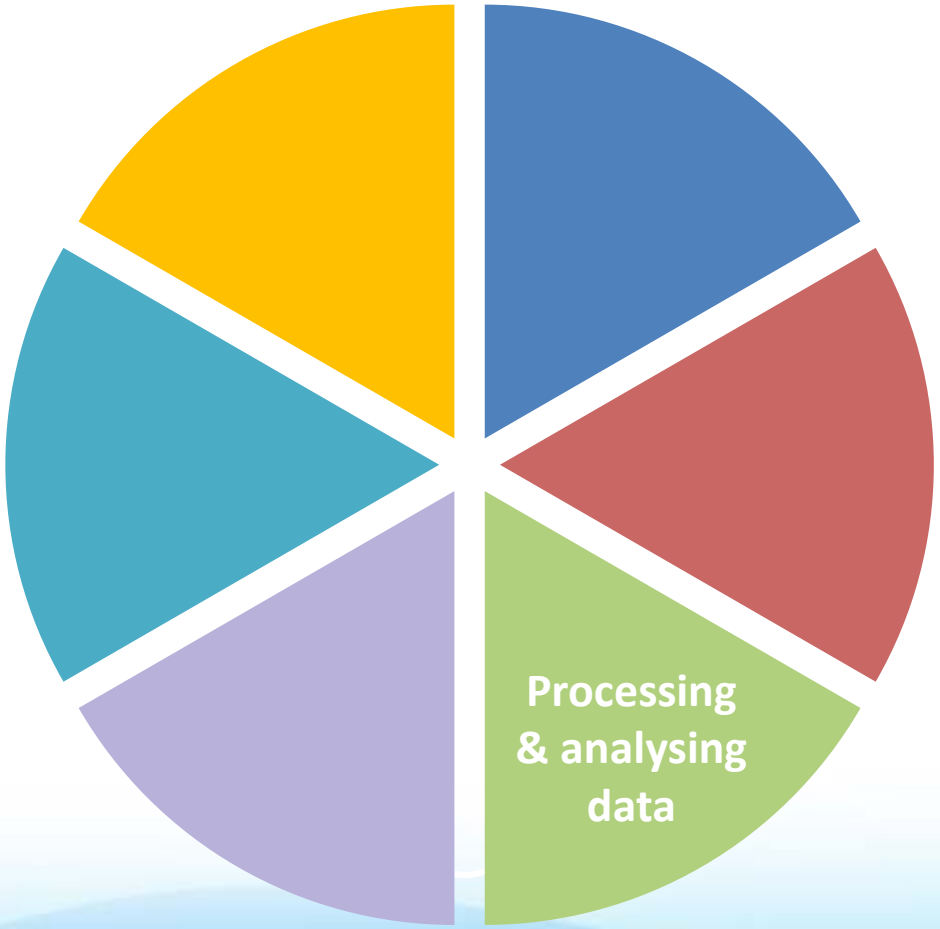
Marineregions.org
a standard for georeferenced marine names

Marine Gazetteer Placedetails

MRGID	http://marineregions.org/mrgid/2401		
Status	Proposed standard		
Name	**Language**	**Name**	**Name source**
	English	Baltic Sea	(1953). Limits of oceans and seas. 3rd edition. IHO Special Publication, 23. International Hydrographic Organization (IHO): Monaco. 38 pp. (look up in IMIS)
PlaceType	IHO Sea Area		
Latitude	58° 56' 39.6" N (58.94434°)		
Longitude	20° 8' 1" E (20.13361°)		

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean



Data curation -
File organization -

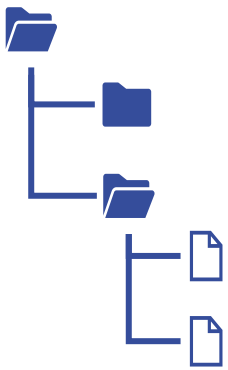
Processing & analysing data

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Processing & analysing

File naming conventions



Recommendations:

Be consistent

Avoid words like ‘draft’, ‘final’... – use version numbers instead (v01, v02)

Use standards (e.g. YYYYMMDD)

Do not use special characters or spaces

...

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

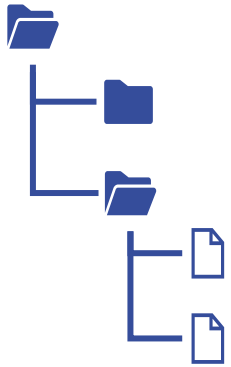
Processing & analysing

File naming conventions

Recommendations:

- Be brief and meaningful
- Not to short; not too long
- Standardise numbers: dates, version iterators
- Avoid ‘draft’, ‘final’... – use version numbers
- Do not use special characters or spaces

- Avoid ambiguous folder names
- Group by similarity, topic, function
- Separate past and active work – create archive folders
- Keep raw and processed apart



Examples of files without a naming convention:	Examples of files with a naming convention:
Meeting notes jan 10.doc	20230110_OT_ODM_exercise1_v01.doc
Third_test.xls	20230110_OT_ODM_exercise1_v03.doc
ProjectProposalFirstVersion.doc	20230109_OT_ODM_EvaluationResults.xls
Project-data.xls	20230109_OT_ODM_RDLC.jpg

VLIZ

Data curation

Document, document, document

Standardise, standardise, standardise

Name	Phone	Birth date	Country
John Smith	445-881-4478	August 12, 1989	Belgium
Fitch, Marie	(876)546-8165	June 15, 72	US
Deere, Alan	+1-189-456-4513	11/12/1965	USA

→

Name	Phone	Birth date	Country
John Smith	445-881-4478	1989-08-12	Belgium
Marie Fitch	876-546-8165	1972-06-15	USA
Alan Deere	189-456-4513	1965-11-12	USA

Keep raw data intact

Document transformation

Version Control

Document Quality Control procedures

Use Open formats

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

- Data archiving

Preserving & archiving data

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

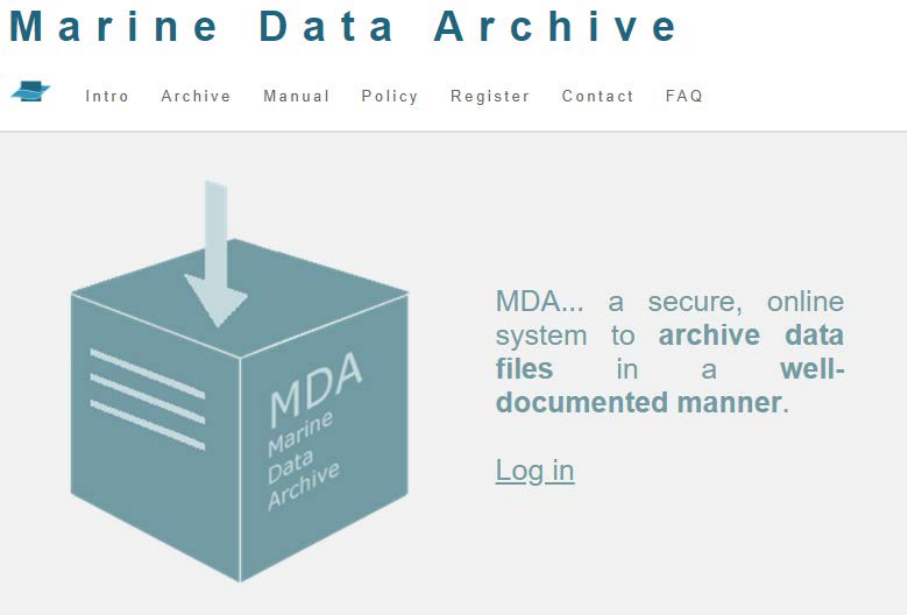
Preserving & archiving

Data archiving

Marine Data Archive - MDA

= trusted data repository for marine, coastal and estuarine research

- **Closed** repository for personal files & projects / collaboration
- **Open** repository for data publication



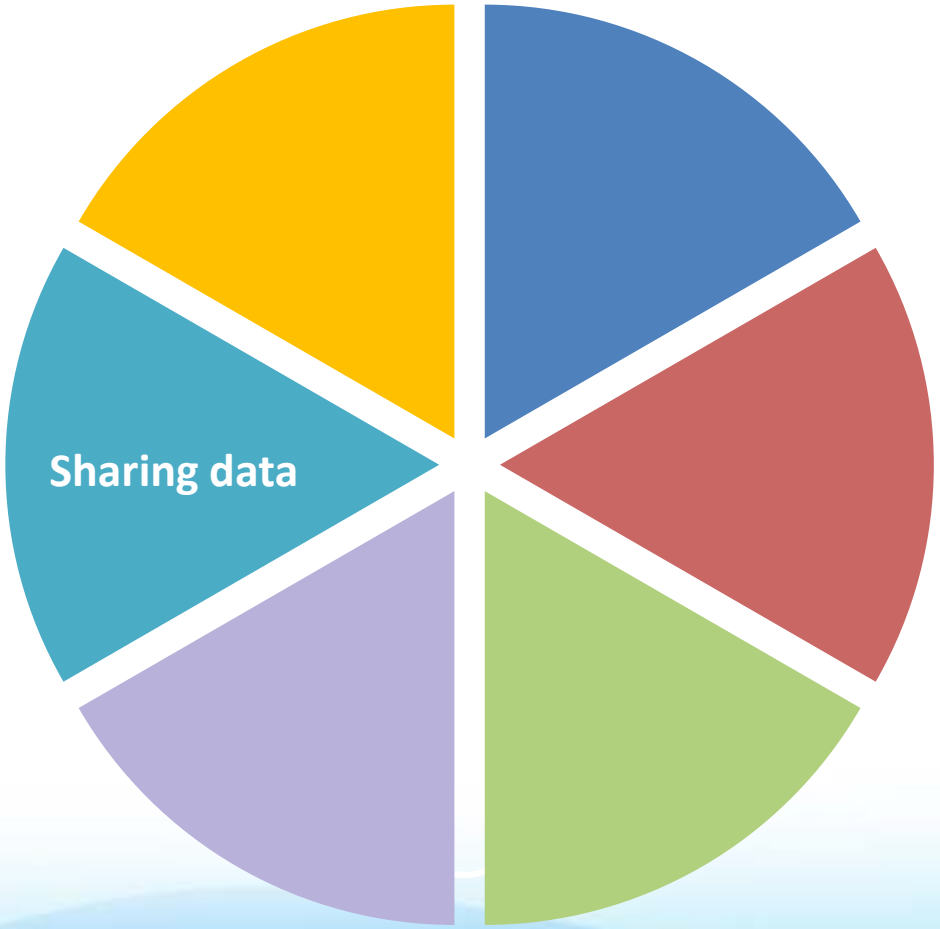
https://mda.vliz.be/

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

- Searchable resources
- IMIS

Sharing data



VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Sharing

Searchable resources

Repositories

- Archiving and sharing
- Generic, discipline specific or institutional

Catalogue

- **Description** (rich metadata) **of and link to data**

Portal

- **Archiving and sharing + interactive tools** (visualisation, combining data, ...)
- **Often thematic**

LifeWatch BELGIUM

med OBIS Ocean Biodiversity Information System

hcmr ΕΛΚΕΘΕ

MARINE FO

EMODnet

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

FAIRsharing.org standards, databases, policies

SeaDataNet

VLIZ

When publishing data

Do remember:

- ORCiD
- Link data to publication and publication to data
- CC 0 or CC BY
- Embargo bad ☹ Release embargo good ☺

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

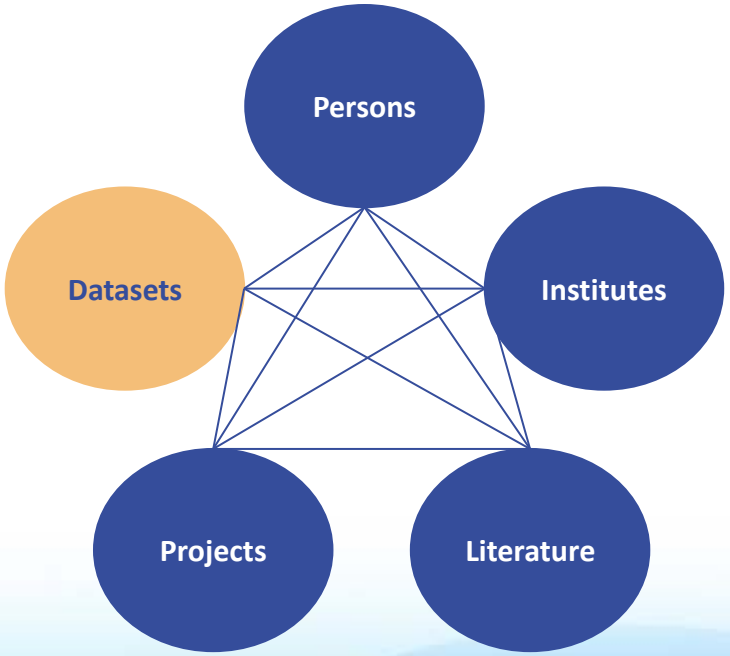
Sharing

IMIS – MarineInfo

MARINE FO

= catalogue with metadata information about:

- All datasets (open / not open)
- Related to marine and coastal research / topics
- Link to data or contact person



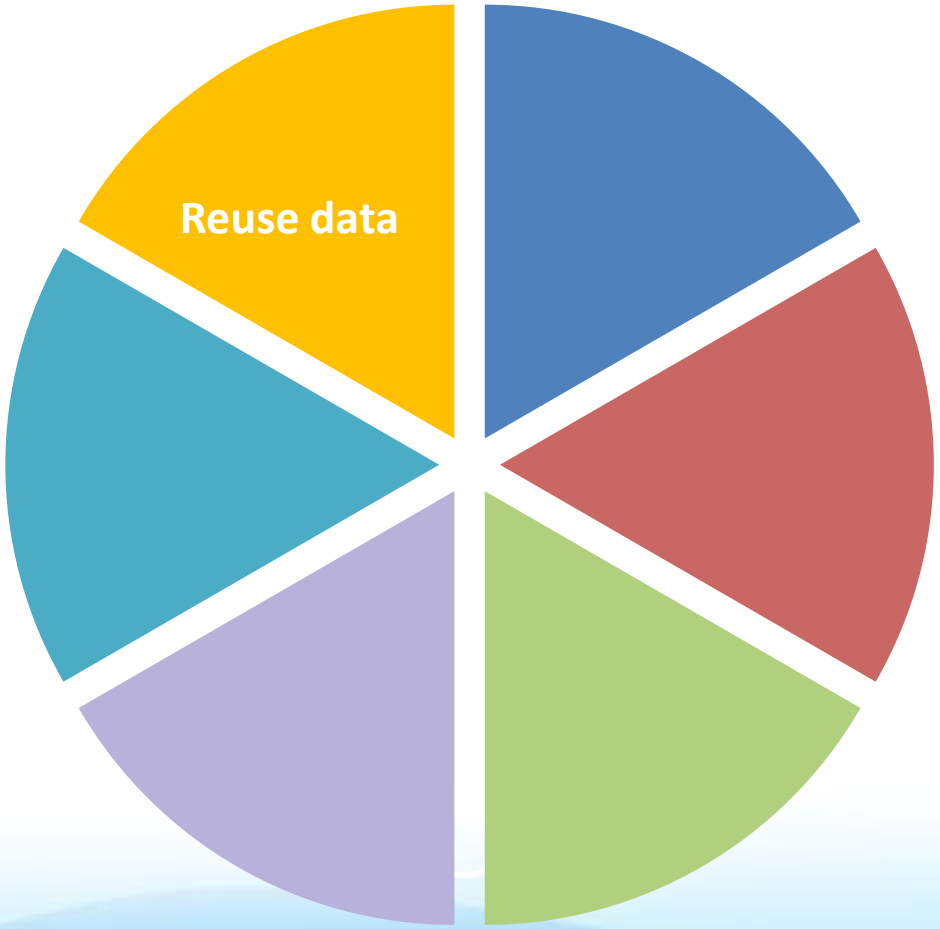
https://www.vliz.be/imis

VLIZ

DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Reusing data

- Provenance
- Rights



VLIZ

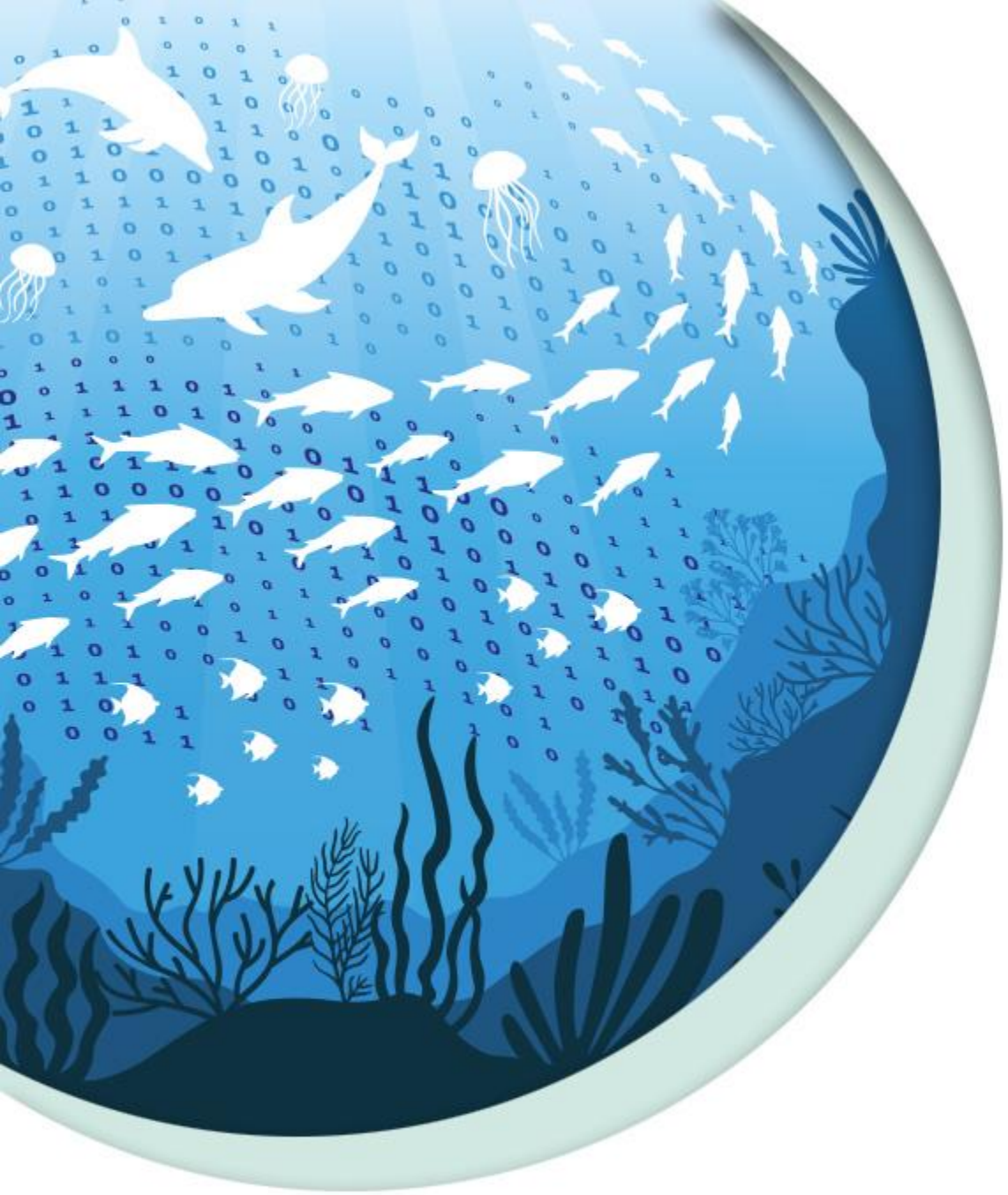
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Reusing data

Provenance and documentation

Usage license and credit

creative commons



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!