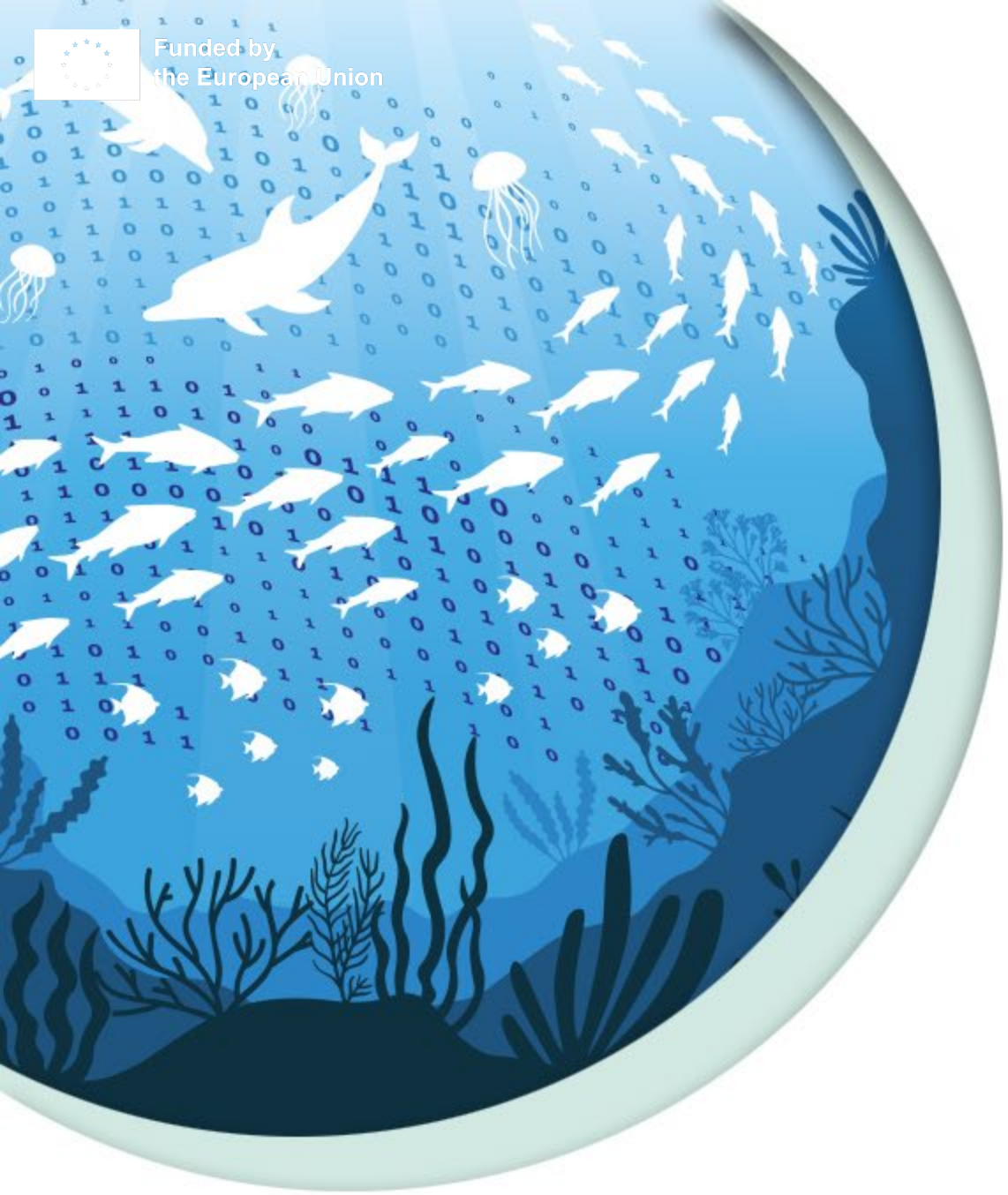
Funded by
the European Union



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

FAIR data, why and how

including general overview of data standards used in EMODnet Biology

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

Research data management

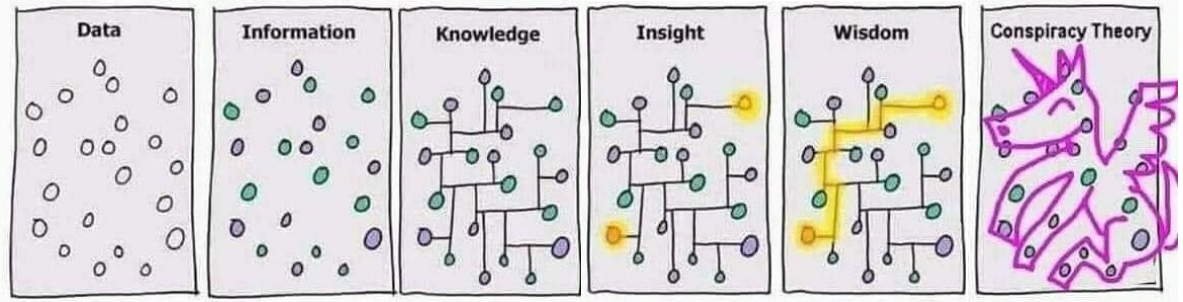
Why it is important

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

RDM

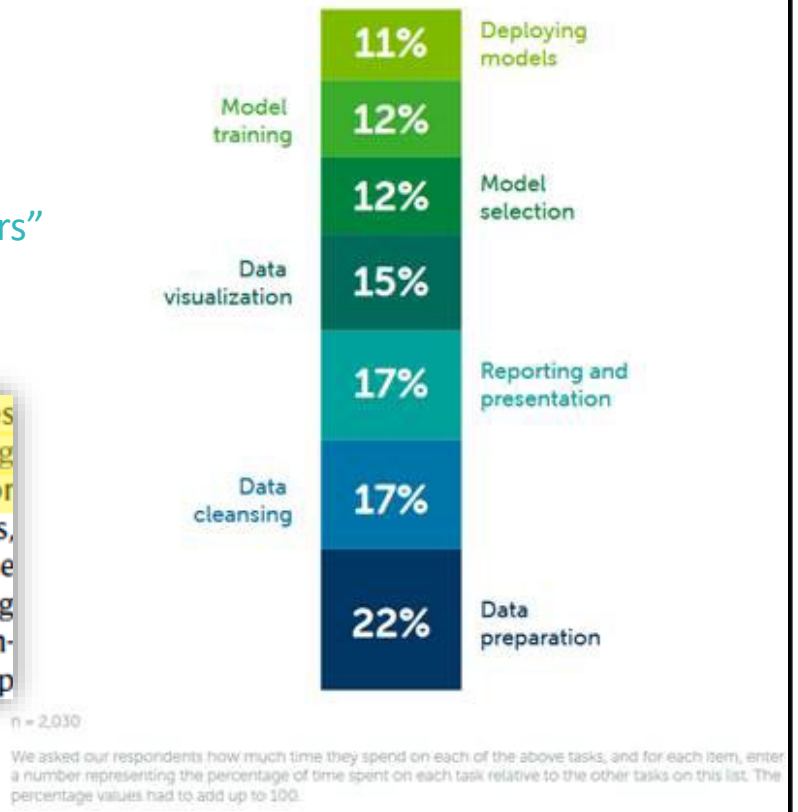
Why it is important

Order, chaos or organized chaos?

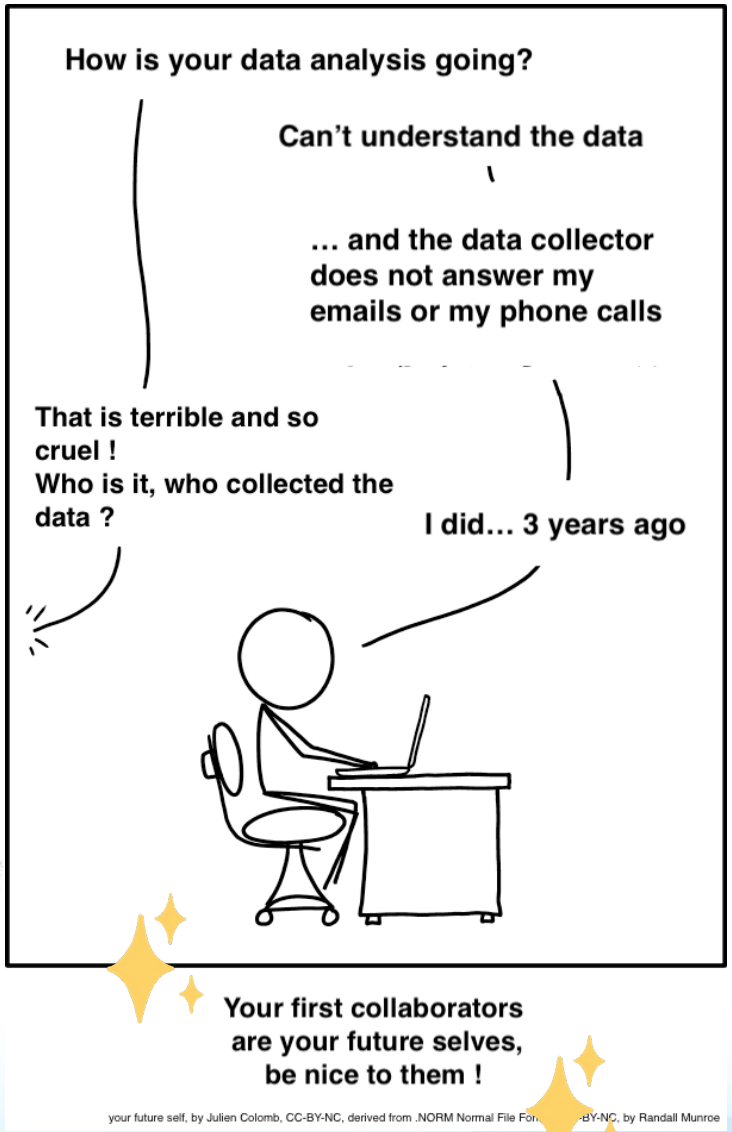
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

"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



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

Personal benefits

Work more efficient & organised

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

RDM

Why it is important

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-BioF
Integration of biodiversity
data into the Digital Twin O
toring

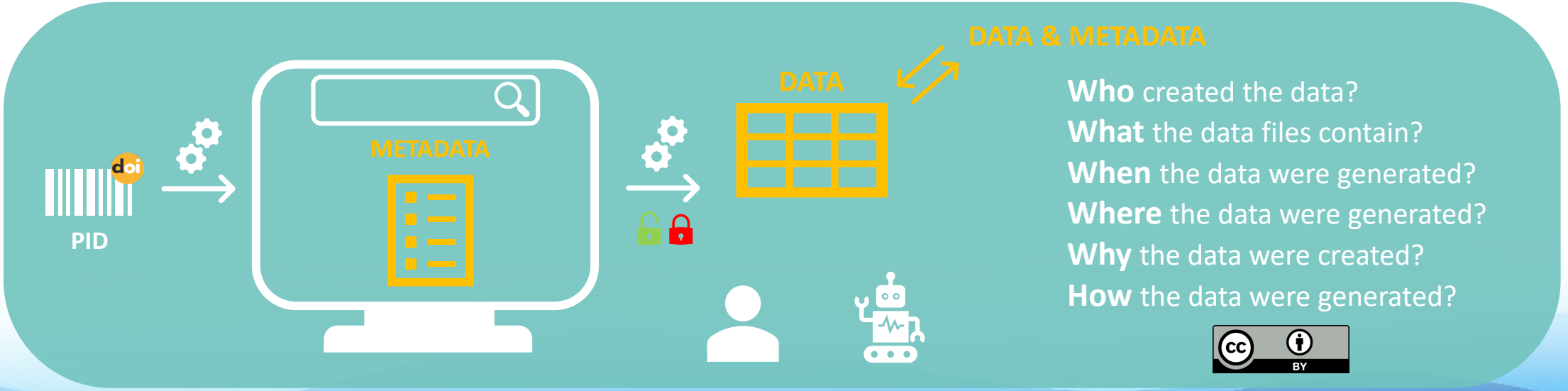
Findable Accessible Interoperable Reusable

- 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

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

The circle of life
Research data management

Why it is important

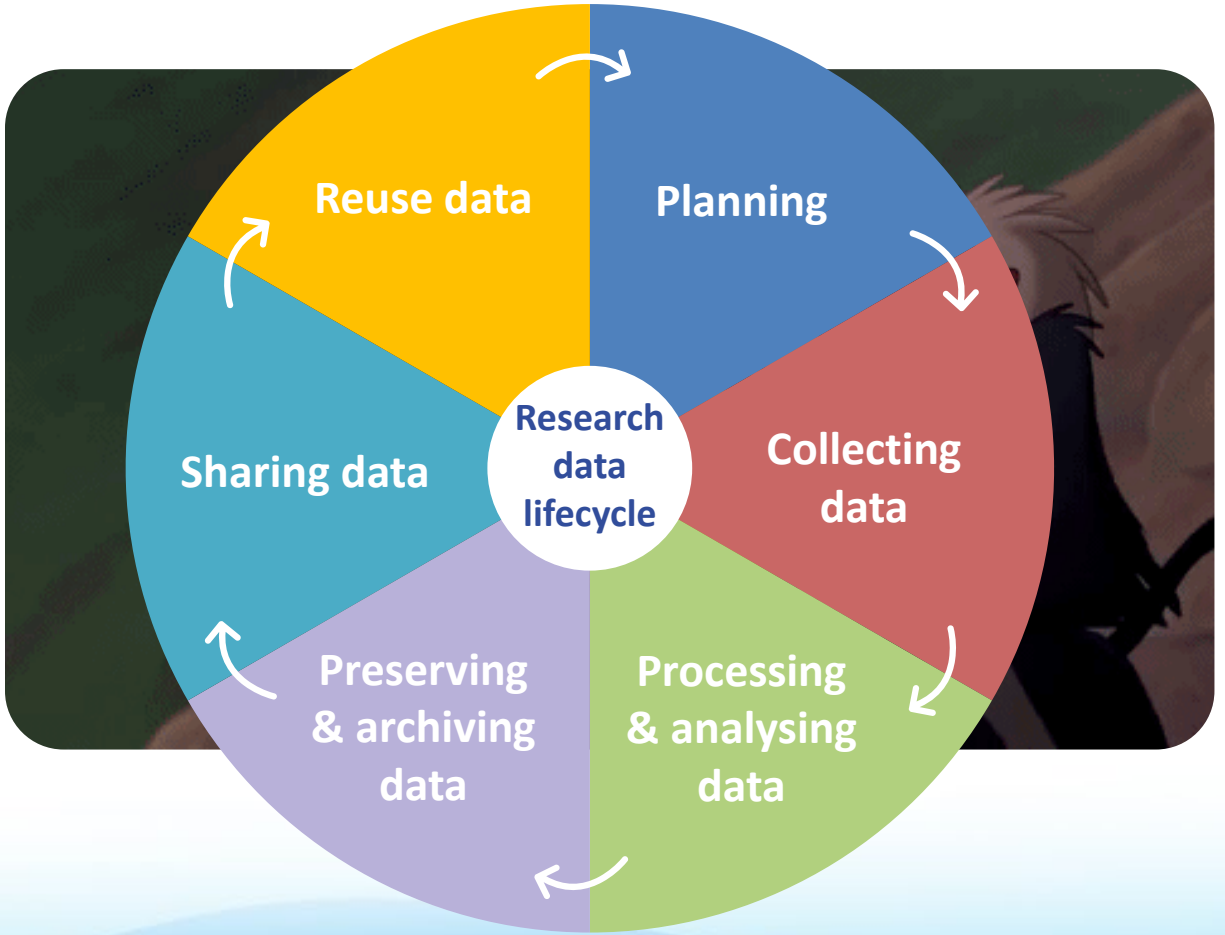
FAIR & Open data

Research data cycle

VLIZ

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

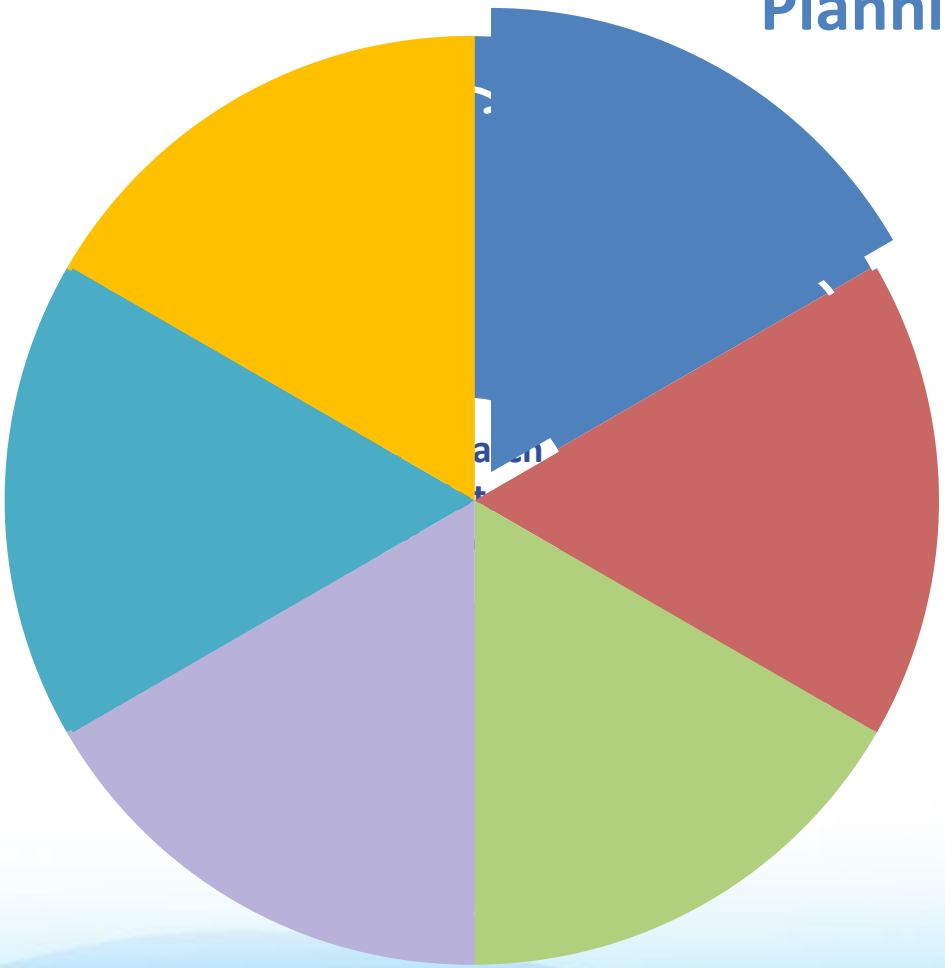
The circle of ~~life~~ data



VLIZ

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

RDM in practice!



VLIZ

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

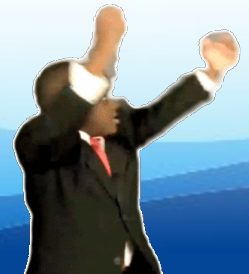
Planning

Data Management Plan

Content of DMP

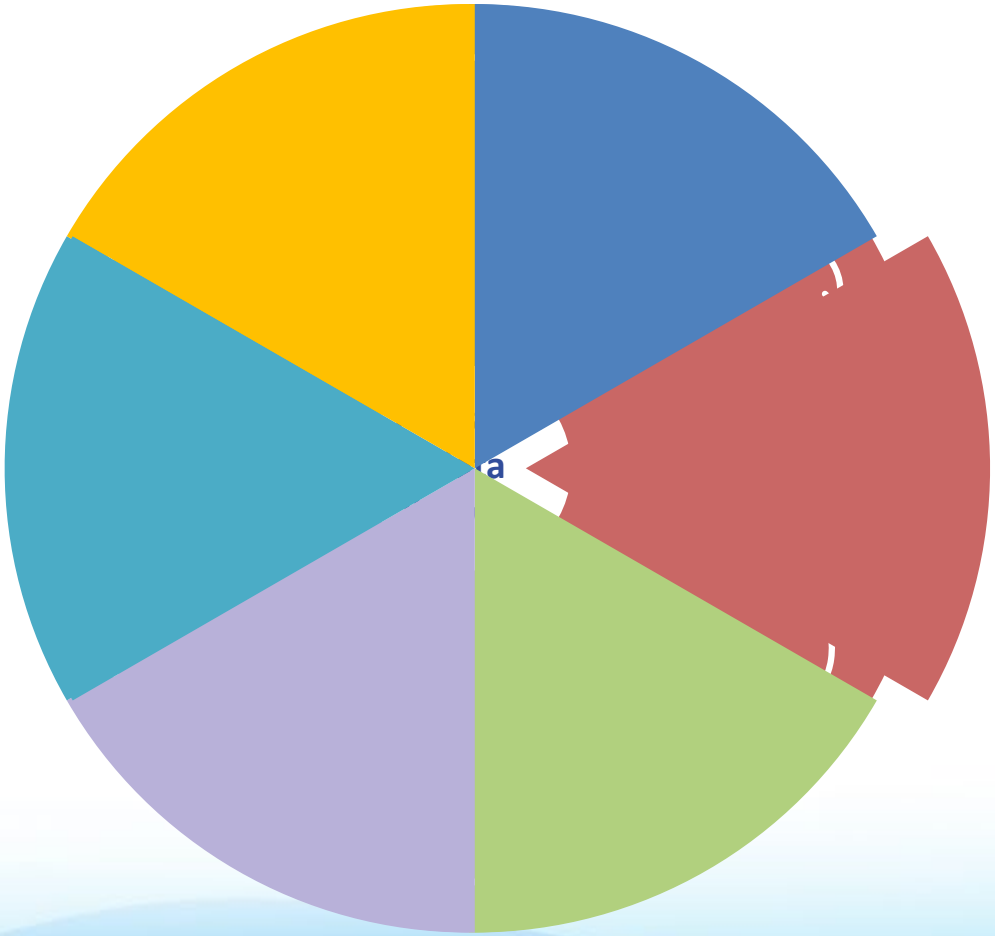
DMP templates

DMPonline.be



VLIZ

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



Collecting data

- Metadata
- Controlled vocabularies
- Standards

VLIZ

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

Collecting

Metadata and Documentation

On three levels

- Project (directory structure)
- Files (naming conventions, READme)
- Data standards & vocabularies

VLIZ

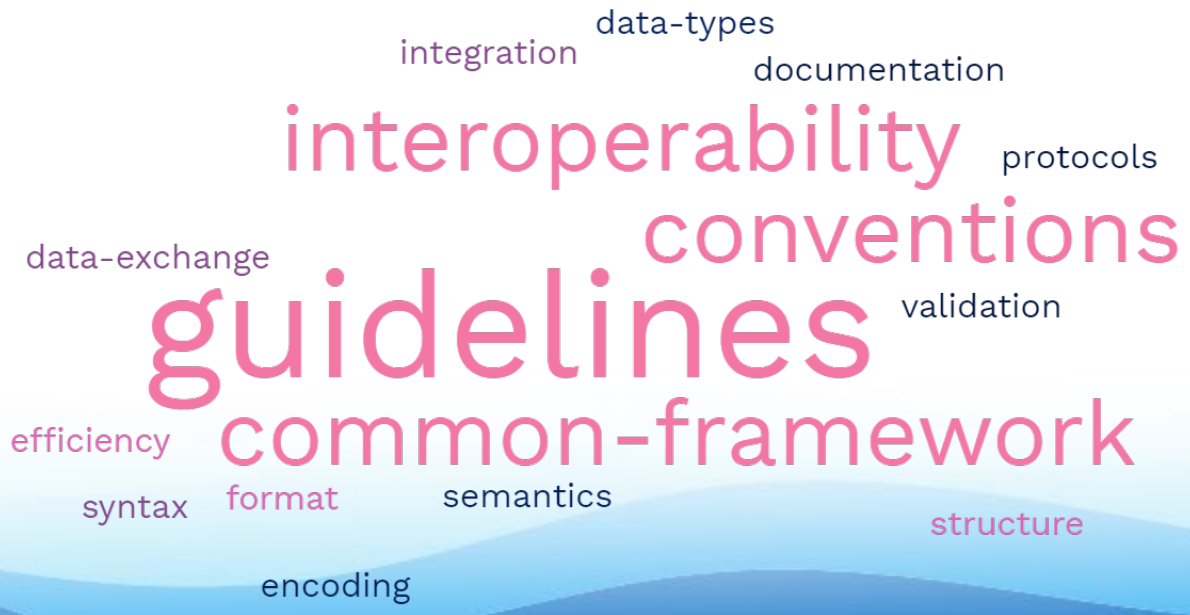
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Collecting

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"



ISO

International Organization for Standardisation

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

Collecting

Data standards

Global & multidisciplinary standards:

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

Example Marine Sciences

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

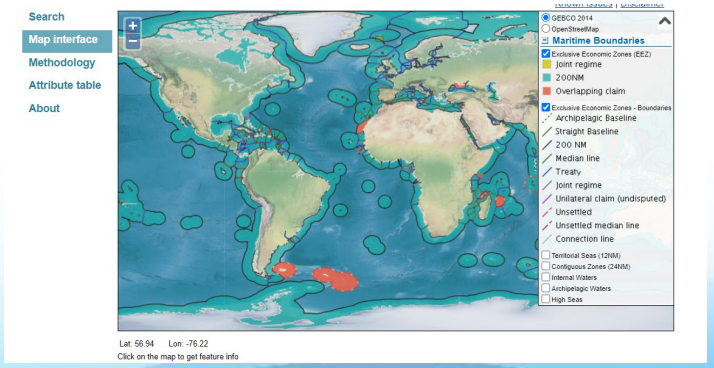
Glaucus atlanticus Forster, 1777

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

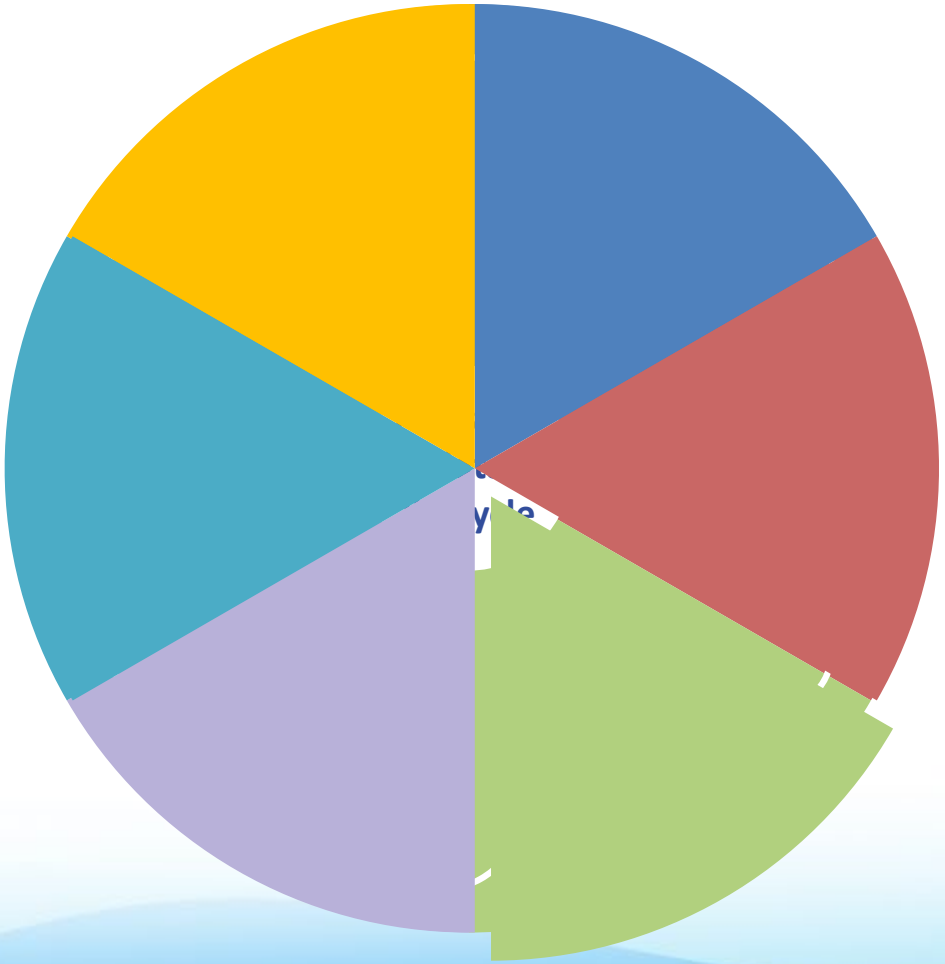
Standard list of marine georeferenced place names & areas

Marineregions.org
a standard for georeferenced marine names



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

Data curation

Processing & analysing

Curation steps

Data exploration

Data transformation

Data enrichment

Data validation

Reproducible procedures

Keep raw data intact

Document transformation

Version Control

Document Quality Control procedures

Use Open formats




VLIZ

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

Processing & analysing

Data curation

✗

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

→

✓

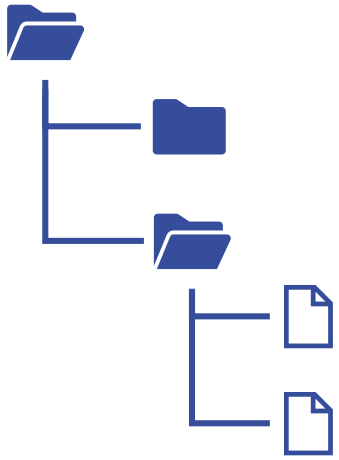
Name	Phone	Birth date	Country
John Smith	445-881-4478	1989-08-12	Belgium
Marie Fitch	876-546-8165	1972-06-15	USA
Alan Deere	189-456-4513	1965-11-12	USA

VLIZ

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

Processing & analysing

File naming conventions



Recommendations:

Be consistent

Use standards (e.g. YYYYMMDD)

Do not use special characters or spaces

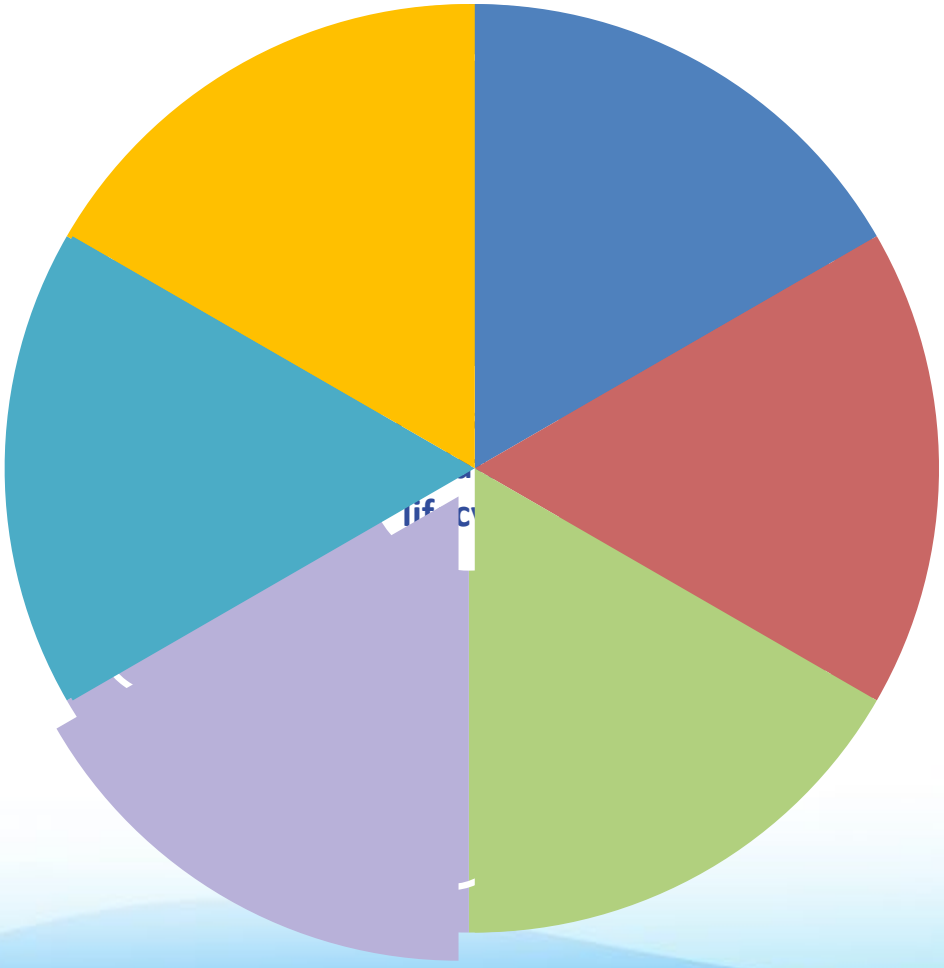
Avoid words like 'draft', 'final'... – use version numbers instead (v01, v02)

...

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

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

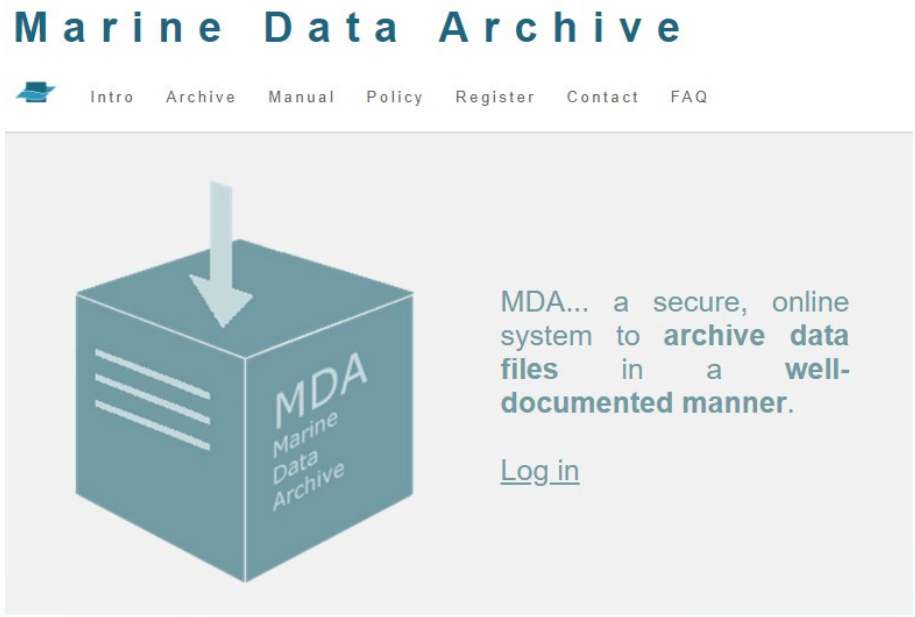
Preserving & archiving

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



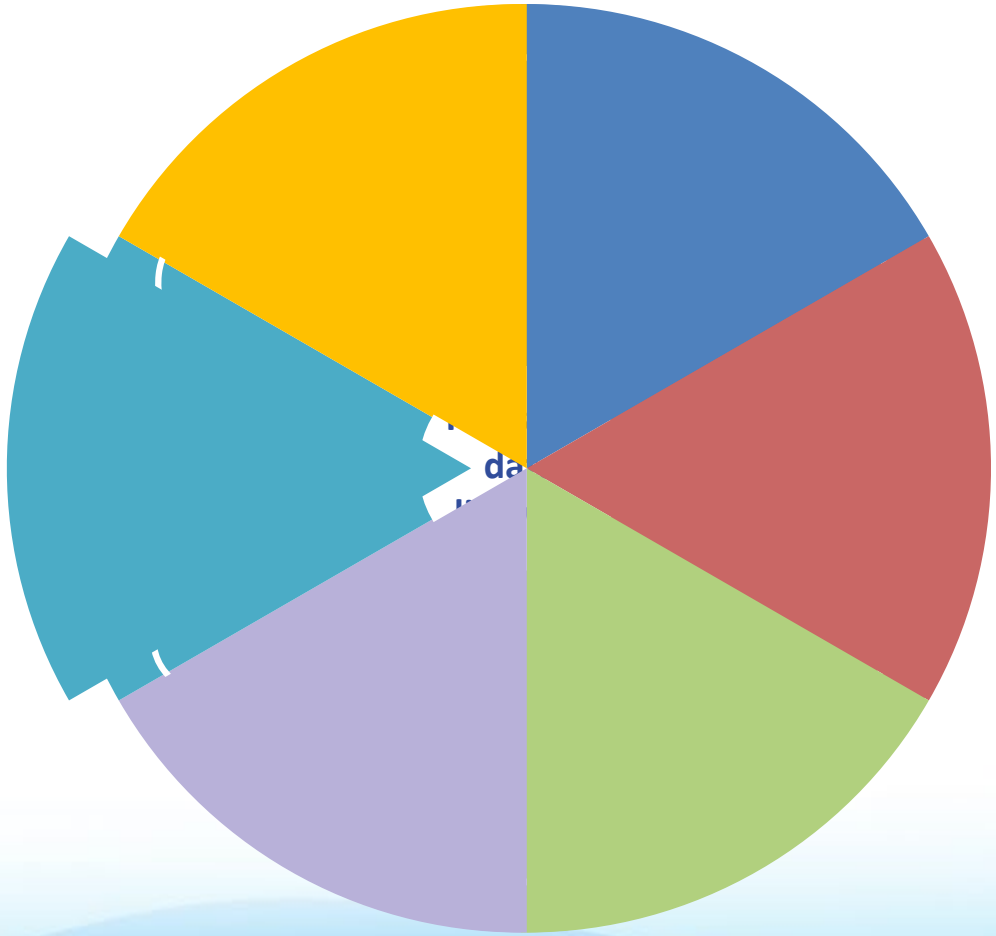
https://mda.vliz.be/

VLIZ

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

- Searchable resources
- IMIS

Sharing data



VLIZ

Searchable resources

Repositories

- **Archiving and sharing**
- **Generic, discipline specific or institutional**

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Catalogue

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

FAIRsharing.org
standards, databases, policies

Portal

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

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

Sharing

Searchable resources

EMODnet

LifeWatch BELGIUM

hcmr ΕΛΚΕΘΕ

med OBIS
Ocean Biodiversity Information System

and more ...

VLIZ
IMIS

SeaDataNet

museum
Belgian Marine Data Centre

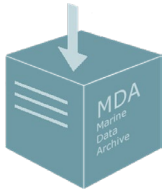
VLIZ

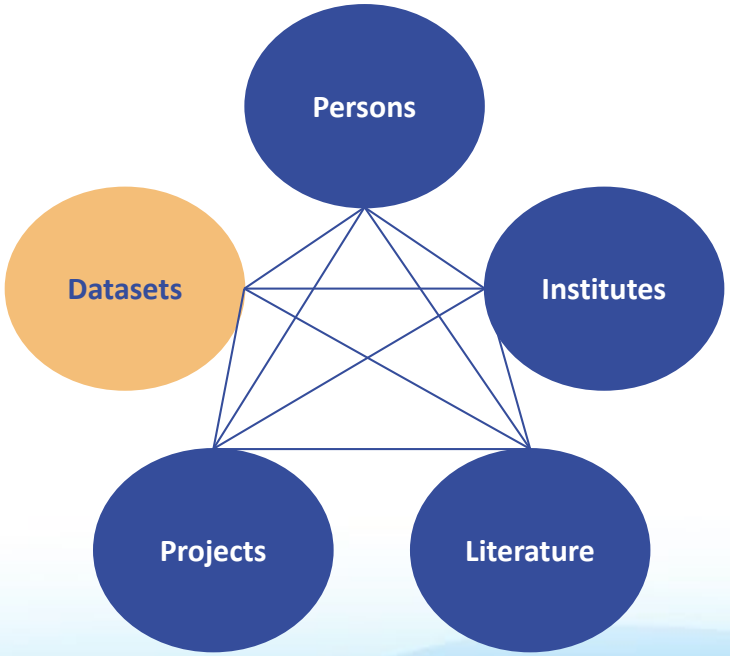
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Sharing

Integrated Marine Information System

= catalogue with metadata information about:

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



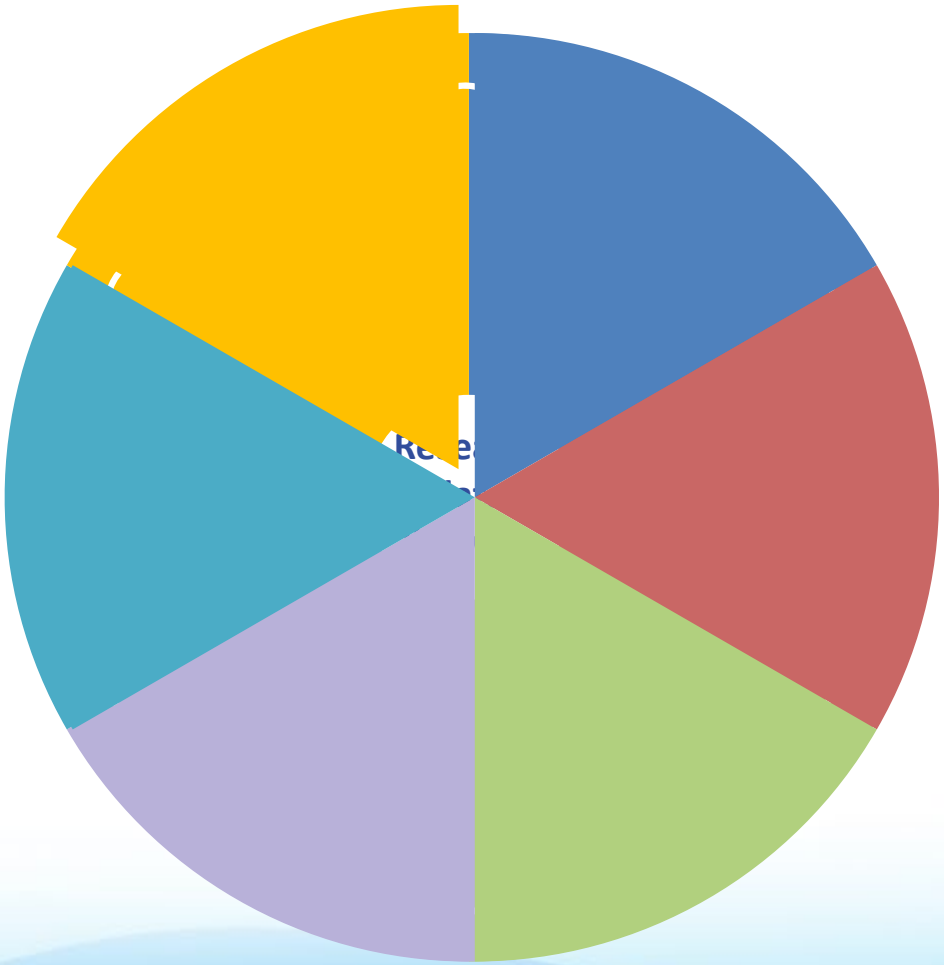
https://www.vliz.be/imis

VLIZ

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

Reusing data

- Provenance
- Rights



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

Reusing data

Provenance and documentation

Usage license and credit

creative commons

VLIZ

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

Data reuse

IMIS

log in

Publications | Institutes | Persons | Datasets | Projects | Maps

[report an error in this record]

LifeWatch observatory data: zooplankton observations by imaging (ZooScan) in the Belgian Part of the North Sea

Citable as data publication

Flanders Marine Institute (VLIZ), Belgium (2023): LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea. https://doi.org/10.14284/584

Download Data

Previous versions (6) view

Contact: data@vliz.be

Access data

EMODnet | EurOBIS | GBIF | OBIS

Archived data

IPT

Also accessible through:
http://rshiny.lifewatch.be/zooscan-data/

Availability:

Notes: Images are available upon request via LifeWatch Belgium (info@lifewatch.be)

Description

In the framework of the Lifewatch marine observatory a number of fixed stations on the Belgian Part of the North Sea (BPNS) are visited on a monthly or seasonal basis using the RV Simon Stevin. A grid of nine stations covers the coastal zone and are sampled monthly. Eight additional stations, located further at sea, are sampled on a seasonal basis. This dataset contains zooplankton observations in the Belgian Part of the North Sea (BPNS) since 2012. Zooplankton is sampled by vertical WP2 net tows, samples scanned with ZooScanner and identification with plankton analyser software, followed by manual validation.

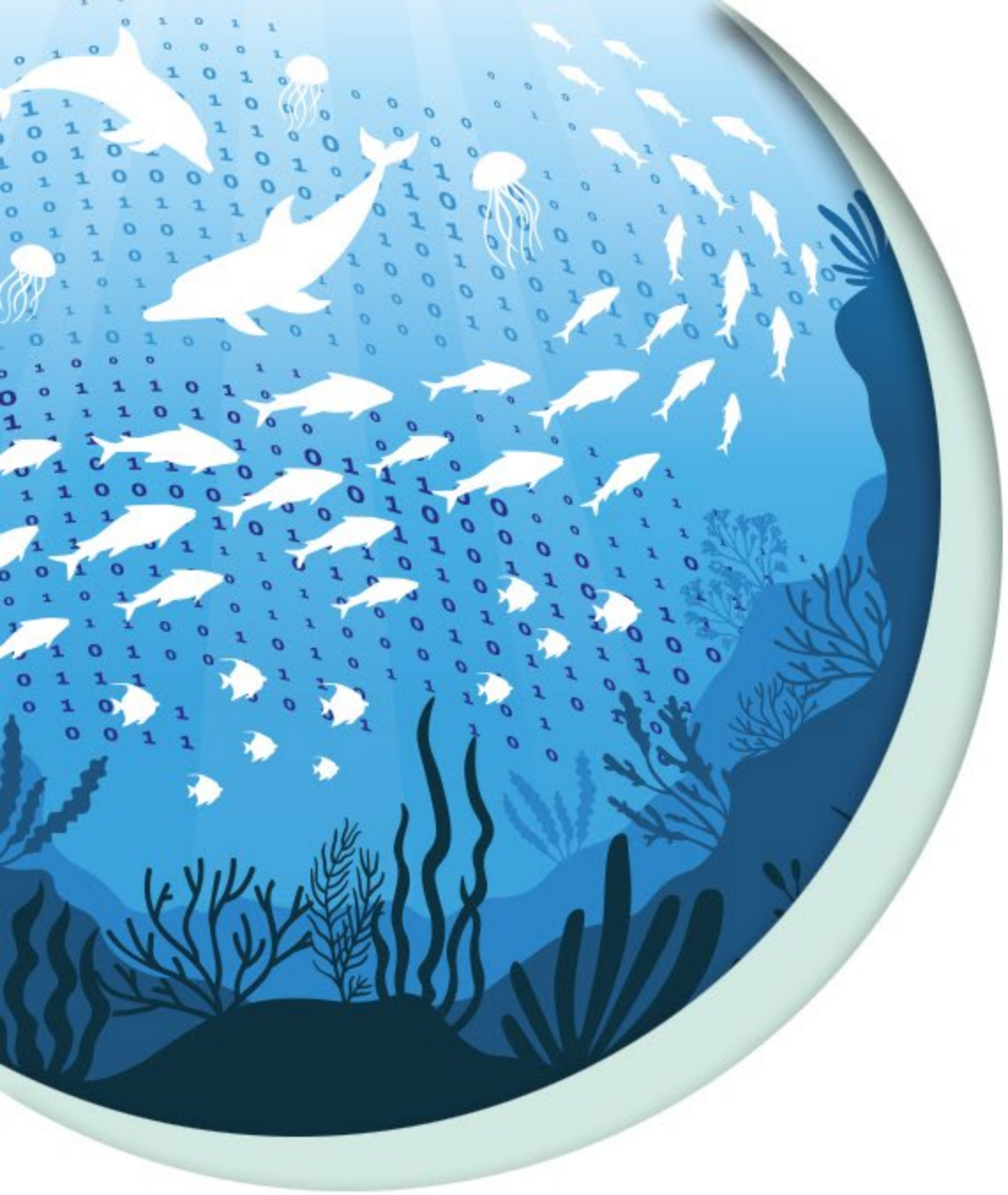
Example IMIS metadata record for dataset: https://vliz.be/en/imis?module=dataset&dasid=4687

VLIZ

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

Relevant sources

- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
- OTGA course: Contributing datasets to EMODnet Biology: https://classroom.oceanteacher.org/course/view.php?id=958



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!