Funded by
the European Union

DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

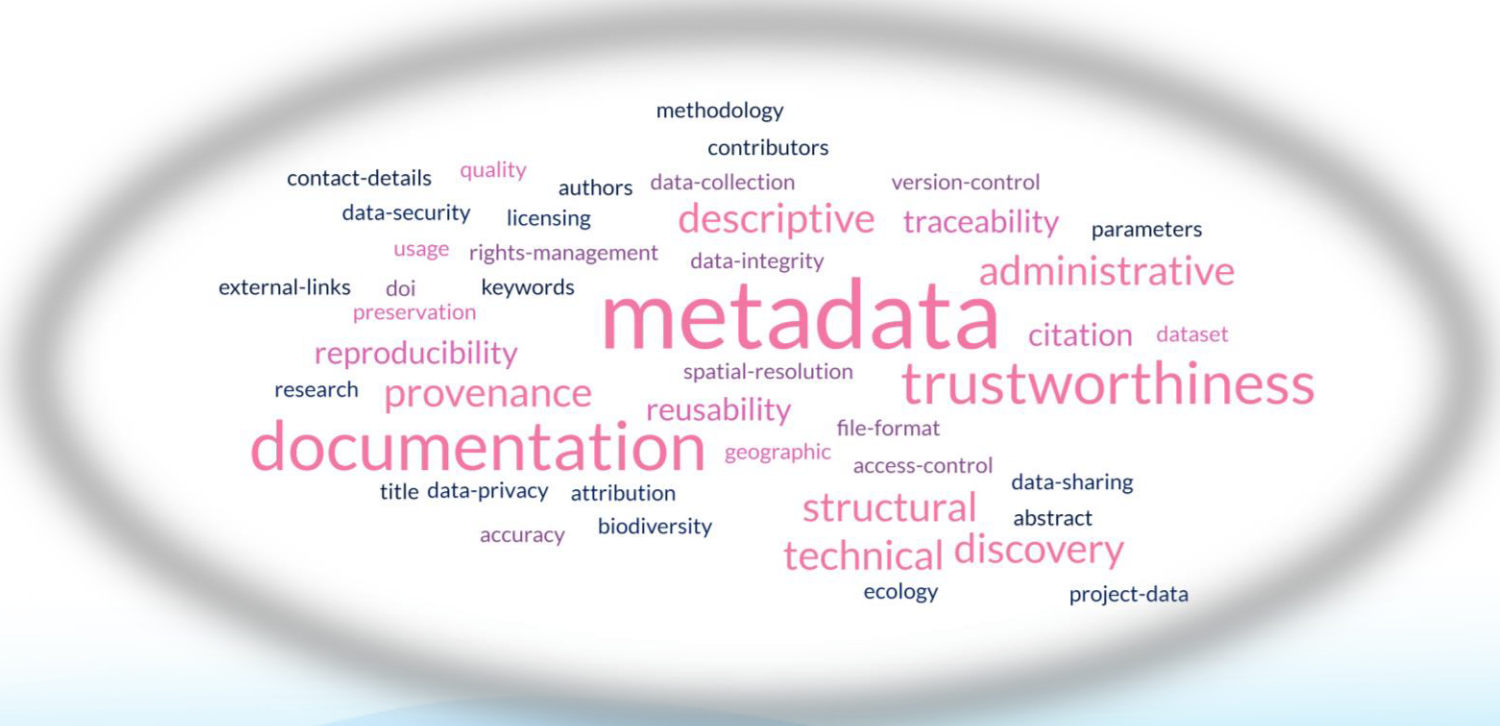
DTO-BioFlow data training workshop:

Metadata

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

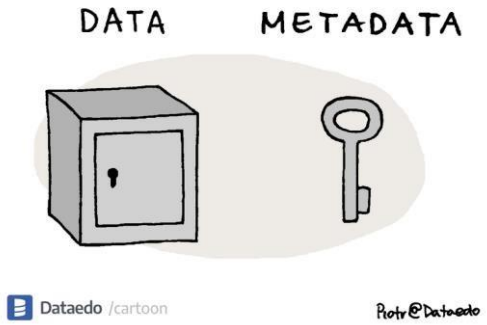
Concept

“Metadata serves as a roadmap to guide users to effectively and efficiently find, understand, use and manage data”

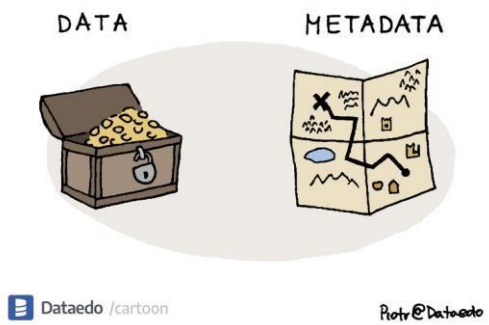


DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

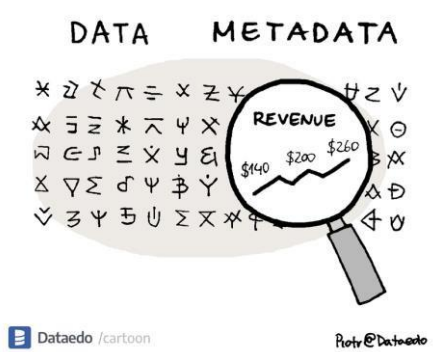
Meta are data that describe other data



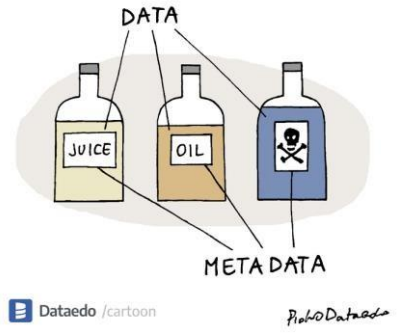
Metadata is the key to data



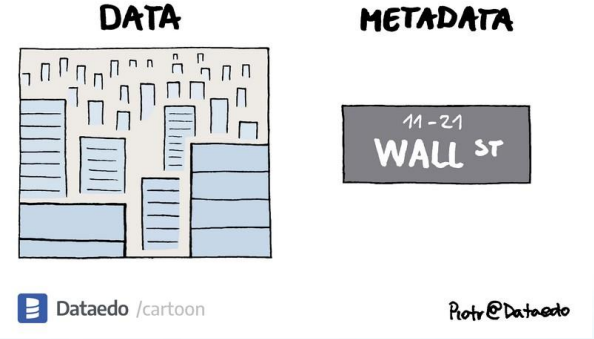
Data is treasure, metadata is a map



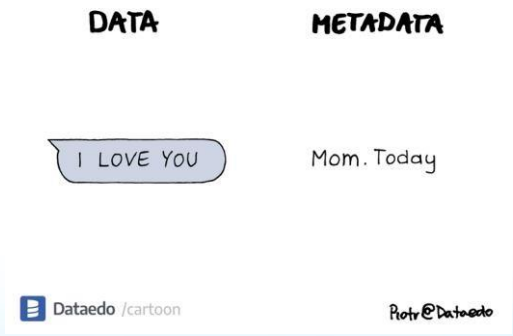
Decoding information



A Matter of Life and Death



How to get to the data?



Context and Source matters

Images source: https://dataedo.com/blog/data-vs-metadata

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

Meta are data that describe other data

There are many different types of metadata – it is not a fixed *thing*: it is defined by what it *does*

Metadata can be used to do the following

- Describe the contents of a data file, e.g. for a spreadsheet you can use metadata to define the column titles, give the units used, define the acronyms
- Describe a set of files, e.g. for a set of images produced by an instrument metadata can provided, for each file, what the file contains, how and when it was created, file location and name, size, version, technical details…
- Provide the provenance information of a data file/set: who, what, how, where, and when it was created and processed
- Provide data discovery and description, e.g. as in data catalogues

Images source: https://dataedo.com/blog/data-vs-metadata

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

Dataset description metadata

The metadata describing an individual file (spreadsheet, text file, and similar) can include

- The meaning of each column title
- The meaning of acronyms, codes, etc
- The data types for each column (string, float, boolean…)
- For each column title, the term from a controlled vocabulary that defines that column
- The units (and the term from a controlled vocabulary for that unit)

Images source: https://dataedo.com/blog/data-vs-metadata

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

Metadata to describe your own spreadsheets

	A	B	C	D	E	F	G	H	I
1	source_mat_id_orig	source_mat_id	chlorophyll	chlorophyll_met	sea_surf_temp	sea_surf_temp_	sea_subsurf_ten	sea_subsurf_ten	sea_surf_salinity	s
2	EMOBON 210701 330 WaSOP1 3um R1	EMOBON_BPNS_V	NA	HPLC	16.848	CTD	16.817	CTD	33.088	C
3	EMOBON 210701 330 WaSOP1 3um R2	EMOBON_BPNS_V	NA	HPLC	16.848	CTD	16.817	CTD	33.088	C
4	EMOBON 210701 330 WaSOP1 3um R3	EMOBON_BPNS_V	NA	HPLC	16.848	CTD	16.817	CTD	33.088	C

	A		C	D	E	F	G	H
1	source_mat_id_orig		tax_id	scientific_name	investigation_type	env_material	collection_date	sampling_event
2	EMOBON 210701 330 WaSOP1 3um R1	-200um	1874687	marine plankton metagenome	metagenome	seawater [ENVO:00002149]; coa	2021-07-01	BPNS_Wa_210701
3	EMOBON 210701 330 WaSOP1 3um R2	-200um	1874687	marine plankton metagenome	metagenome	seawater [ENVO:00002149]; coa	2021-07-01	BPNS_Wa_210701
4	EMOBON 210701 330 WaSOP1 3um R3	-200um	1874687	marine plankton metagenome	metagenome	seawater [ENVO:00002149]; coa	2021-07-01	BPNS_Wa_210701
5	EMOBON 210701 330 WaSOP1 3um R4	-200um	1874687	marine plankton metagenome	metagenome	seawater [ENVO:00002149]; coa	2021-07-01	BPNS_Wa_210701

	A	B	C	D	E
1	LogsheetColumnTitle	DataTypeIn	Observable_property_url	Unit	Unit_URL
2	alkalinity	xsd:float	http://vocab.nerc.ac.uk/collection/P01/current/ALKYZZXX/	mEq/l	http://vocab.nerc.ac.uk/collection/P06/current/MEQL/
3	alkalinity_method	xsd:string	NA	NA	
4	ammonium	xsd:float	http://vocab.nerc.ac.uk/collection/P35/current/EPC00009/	umol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
5	ammonium_method	xsd:string	NA	NA	
6	bac_prod	xsd:float	http://vocab.nerc.ac.uk/collection/P02/current/UPTH/	ug/m3/d	https://data.emobon.embrc.eu/ns/unit_vocab#"Micrograms
7	bac_prod_method	xsd:string	NA	NA	
8	biomass	xsd:list	https://data.emobon.embrc.eu/ns/observableproperty_vocab#Bior	g/l	http://vocab.nerc.ac.uk/collection/P06/current/GPRL/
9	biomass_method	xsd:list	NA	NA	

edo.com/blog/data-vs-metadata

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

Metadata for DwC-A files

occurrenceID	env_broad_scale	env_local_scale	env_medium_scal	DNA_Sequence	SOP	target_gene	pcr_primer_forward
ARMS_BelgiumCo	marine benthic bic	sandy sediment [E	marine reef biome	TGGTGGAGTGATT	https://github.com/arms-mbon/docur	18S	GGWACWGGWTGAACWG
ARMS_BelgiumCo	marine benthic bic	sandy sediment [E	marine reef biome	TGGTGGAGCGATT	https://github.com/arms-mbon/docur	18S	GGWACWGGWTGAACWG
ARMS_BelgiumCo	marine benthic bic	sandy sediment [E	marine reef biome	TGGTGGAGCGATT	https://github.com/arms-mbon/docur	18S	GGWACWGGWTGAACWG
ARMS_BelgiumCo	marine benthic bic	sandy sediment [E	marine reef biome	TGGTGGAGTGATT	https://github.com/arms-mbon/docur	18S	GGWACWGGWTGAACWG

	occurrenceID	basisOfRecord	institutionCode	datasetName	occurrenceStatus	eventID	materialSampleID	eventDate	o
2	ARMS_BelgiumCoast	MaterialSample	ARMS-MBON	data_release_001	present	ARMS_BelgiumCoas	ARMS_BelgiumCoast_AZBE1_2019	2019-09-24/2020-	
3	ARMS_BelgiumCoast	MaterialSample	ARMS-MBON	data_release_001	present	ARMS_BelgiumCoas	ARMS_BelgiumCoast_AZBE1_2019	2019-09-24/2020-	
4	ARMS_BelgiumCoast	MaterialSample	ARMS-MBON	data_release_001	present	ARMS_BelgiumCoas	ARMS_BelgiumCoast_AZBE1_2019	2019-09-24/2020-	
5	ARMS_BelgiumCoast	MaterialSample	ARMS-MBON	data_release_001	present	ARMS_BelgiumCoas	ARMS_BelgiumCoast_AZBE1_2019	2019-09-24/2020-	
6	ARMS_BelgiumCoast	MaterialSample	ARMS-MBON	data_release_001	present	ARMS_BelgiumCoas	ARMS_BelgiumCoast_AZBE1_2019	2019-09-24/2020-	

occurrenceID

An identifier for the dwc:Occurrence (as opposed to a particular digital record of the dwc:Occurrence). In absence of a persistent global unique identifier, construct one from a combination of identifiers in the rec that will most closely make the dwc:occurrenceID globally unique. See also https://dwc.tdwg.org/terms/#dwc:occurrenceID

Examples: `http://arctos.database.museum/guid/MSB:Mamm:233627`; `000866d2-c177-4648-a200-ead4007051b9`; `urn:catalog:UWBM:Bird:89776`

basisOfRecord

The specific nature of the data record. See also https://dwc.tdwg.org/terms/#dwc:basisOfRecc

Examples: `MaterialEntity`; `PreservedSpecimen`; `FossilSpecimen`; `LivingSpecimen`; `Materia` `HumanObservation`; `MachineObservation`; `Taxon`; `Occurrence`; `MaterialCitation`

Qualname	http://rs.tdwg.org/dwc/terms/basisOfRecord
Namespace	http://rs.tdwg.org/dwc/terms/
Group	Record-level
Data Type	Vocabulary: http://rs.gbif.org/vocabulary/dwc/basis_of_record_2024-02-19.xml
Required	true

env_broad_scale

In this field, report which major environmental system your sample or specimen came from. The systems identified should have a coarse spatial grain, to provide the general environmental context of where the sampling was done (e.g. were you in the desert or a rainforest?). We recommend using subclasses of ENVO's biome class: http://purl.obolibrary.org/obo/ENVO_00000428. Format (one term): termLabel [termID], Format (multiple terms): termLabel [termID]|termLabel [termID]|termLabel [termID]. Example: Annotating a water sample from the photic zone in middle of the Atlantic Ocean, consider: oceanic epipelagic zone biome [ENVO:01000033]. Example: Annotating a sample from the Amazon rainforest consider: tropical moist broadleaf forest biome [ENVO:01000228]. If needed, request new terms on the ENVO tracker, identified here: http://www.obofoundry.org/ontology/envo.html

Examples: forest biome [ENVO:01000174]

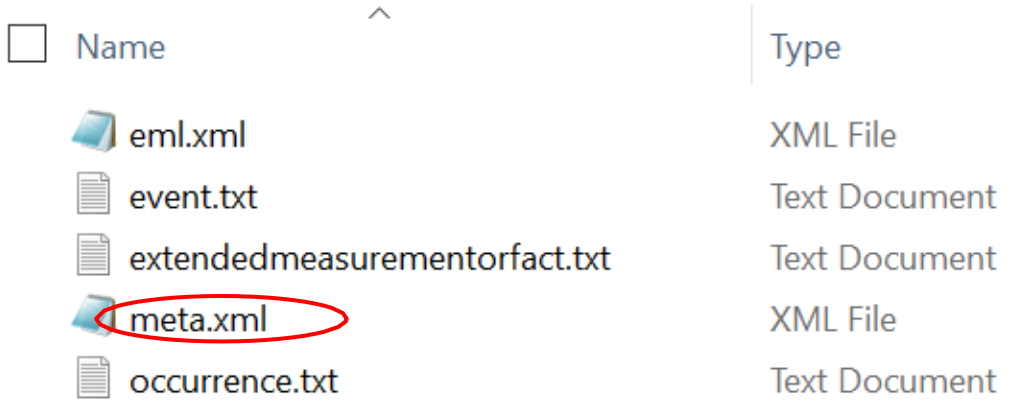
Qualname	https://w3id.org/mixs/0000012
Namespace	https://w3id.org/mixs/
Group	environment
Data Type	
Required	false

env_local_scale

In this field, report the entity or entities which are in your sample or specimen's local vicinity and which you believe have significant causal influences on your sample or specimen. Please use terms that are present in ENVO and which are of smaller spatial grain than your entry for env_broad_scale. Format

Images source: https://dataedo.com/blog/data-vs-metadata

Metadata for DwC-A files



meta.xml - Notepad

File Edit Format View Help

```
<archive xmlns="http://rs.tdwg.org/dwc/text/" metadata="eml.xml">
  <core encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n" fi
    <files>
      <location>event.txt</location>
    </files>
    <id index="0" />
    <field index="1" term="http://purl.org/dc/terms/modified"/>
    <field index="2" term="http://purl.org/dc/terms/language"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/datasetID"/>
    <field index="4" term="http://rs.tdwg.org/dwc/terms/datasetName"/>
    <field index="5" term="http://rs.tdwg.org/dwc/terms/dynamicProperties"
    <field index="6" term="http://rs.tdwg.org/dwc/terms/eventID"/>
    <field index="7" term="http://rs.tdwg.org/dwc/terms/parentEventID"/>
    <field index="8" term="http://rs.tdwg.org/dwc/terms/samplingProtocol"/
    <field index="9" term="http://rs.tdwg.org/dwc/terms/sampleSizeValue"/>
    <field index="10" term="http://rs.tdwg.org/dwc/terms/sampleSizeUnit"/>
    <field index="11" term="http://rs.tdwg.org/dwc/terms/samplingEffort"/>
    <field index="12" term="http://rs.tdwg.org/dwc/terms/eventDate"/>
    <field index="13" term="http://rs.tdwg.org/dwc/terms/eventTime"/>
    <field index="14" term="http://rs.tdwg.org/dwc/terms/startDayOfYear"/>
    <field index="15" term="http://rs.tdwg.org/dwc/terms/endDayOfYear"/>
    <field index="16" term="http://rs.tdwg.org/dwc/terms/year"/>
```

Images source: https://dataedo.com/blog/data-vs-mtadata

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

Dataset description metadata

The metadata describing an individual file (spreadsheet, text file, and similar) can include

- The meaning of each column title
- The meaning of acronyms, codes, etc
- The data types for each column (string, float, boolean…)
- For each column title, the term from a controlled vocabulary that defines that column
- The units (and the term from a controlled vocabulary for that unit)

The metadata describing a set of files can include

- File name, file location, file size, file type
- Link to a sample ID (where that file was created from a sample)
- Date, location, and method of creation
- Technical values (images, videos…)

Images source: https://dataedo.com/blog/data-vs-metadata

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

Metadata to describe a file-set

ObservatoryID	UnitID	EventID	Filename	Filetype	PlateNumber	Position	Download URL
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_Belgium_zbe1_190924 - 200303_001.jp	Image	Not Provided	Not Provided	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_1B.JPG	Image	1	Bottom	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_1B_1.JPG	Image	1	Bottom	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_1T.JPG	Image	1	Top	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_1T_1.JPG	Image	1	Top	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_2B.JPG	Image	2	Bottom	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_2B_1.JPG	Image	2	Bottom	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_2T.JPG	Image	2	Top	https://files.plutof.u
BelgiumCoast	AZBE1	ARMS_BelgiumCoast_AZBE1_20190924_20200303	ARMS_zbe1_190924-200303_IMG_2T_1.JPG	Image	2	Top	https://files.plutof.u

Column title	Definition	Data type	Property	ObservablePropertyUrl	Unit	UnitURL
ObservatoryID	Observatory identifier	xsd:string	schema:identifier	none	none	none
UnitID	ARMS unit identifier	xsd:string	schema:identifier	none	none	none
EventID	Sampling event identifier	xsd:string	schema:identifier	http://rs.tdwg.org/dwc/terms/eventID	none	none
Filename	File name	xsd:string	schema:name	none	none	none
Filetype	File type	xsd:string	schema:fileFormat	none	none	none
PlateNumber	ARMS plate number (0 at bot	xsd:integer	schema:description	none	none	none
Position	ARMS plate face (top/bottom	xsd:string	schema:description	none	none	none
Download URL	URL to download the image	xsd:anyURI	schema:url	none	none	none

Images source: https://dataedo.com/blog/data-vs-metadata

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

Provenance metadata

These metadata should describe

- Who did the work
- Where the work was done (samples collected, samples processed)
- When the work was done (ditto)
- How the work was done (protocols followed, methodologies employed, software and instrument settings)
- What was used (instruments, devices, software, links/IDs for physical samples and data used, links/IDs for physical samples or data created)

And ideally the metadata should provide this information for each activity step carried out

Images source: https://dataedo.com/blog/data-vs-metadata

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

Provenance metadata

```
## The provenance metadata for the sampling activity (Step 1)
<Sampling>
name: water sampling
activity-type: sampling
spatial-coverage: see Belgian EEZ
temporal-coverage: 2021-01-01
agent: see Jane Smith
protocol: see BigProject_WaterSamples
result: see BigProject_belgium_water_10m
device: see NiceMarineStation rosette#1
permit: see ABS permit

## The provenance metadata for the observing activity (Step 2)
<Observing>
name: in-situ measurement of water properties
activity-type: observing
spatial-coverage: see Belgian EEZ
temporal-coverage: 2021-01-01
agent: see Jane Smith
protocol: see CTD_standard_procedure_2
result: http://vocab.nerc.ac.uk/collection/P02/current/TEMP, http://voc
```

```
## Agents
name: Jane Smith
identifier: https://orcid.org/0000-0001-0001-0001
email: jane.smith@somewhere.com
role: field and laboratory technician
organization:
  name: Nice Marine Station
  identifier: https://edmo.seadatanet.org/report/00000
  website: https://NiceMarineStation.eu
  email: info@nicemarinestation.eu

## Locations
name: Belgian EEZ
identifier: http://marineregions.org/mrgid/3293
geography: latitude: 51.249995, longitude: 2.85327

## Samples
identifier: BigProject_belgium_water_10m
description: water sample to determine plankton community and extract DNA
keyword: coastal sea water - http://purl.obolibrary.org/obo/ENVO_00002150, marine plaknton metagenome - NCBI:txid1874687
## Protocols
name: BigProject_WaterSamples
description: Collecting and pre-filtering (removal of larger particles) of water samples to be used later for FlowCam and eDNA work
distribution: https://www.protocols.io/BigProject_WaterSamples.pdf

## Devices
name: NiceMarineStation rosette#1
type: Niskin bottles on a Rosette sampler
platform:
  identifier: IMO:1234567
  type: research vessel - https://vocab.nerc.ac.uk/collection/L05/31
  name:  RV OurBigShip
```

://dataedo.com/blog/data-vs-metadata

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

Provenance in DwC-A files

decimalLatitude	The geographic latitude (in dwc:geodeticDatum) of the
locationID	An identifier for the set of dcterms:Location information. specific to the data set. See also https://dwc.tdwg.org/te
eventDate	The date-time or interval during which a dwc:Event oc dwc:Event was recorded. Not suitable for a time in a d
samplingProtocol	The names of, references to, or descriptions of the methods or protoc https://dwc.tdwg.org/terms/#dwc:samplingProtocol
samp_collec_method	The method employed for collecting the sample
samp_mat_process	Any processing applied to the sample during or after retrieving the sample from environm accepts OBI, for a browser of OBI (v 2018-02-12) terms please see
recordedBy	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original dwc:Occurrence. The primary collector or observer, especially one who applies a personal identifier (dwc:recordNumber), should be listed first. See also https://dwc.tdwg.org/terms/#dwc:recordedBy

Images source: https://dataedo.com/blog/data-vs-metadata

Catalogue metadata

These metadata are to allow data to be discovered and described in a data catalogue

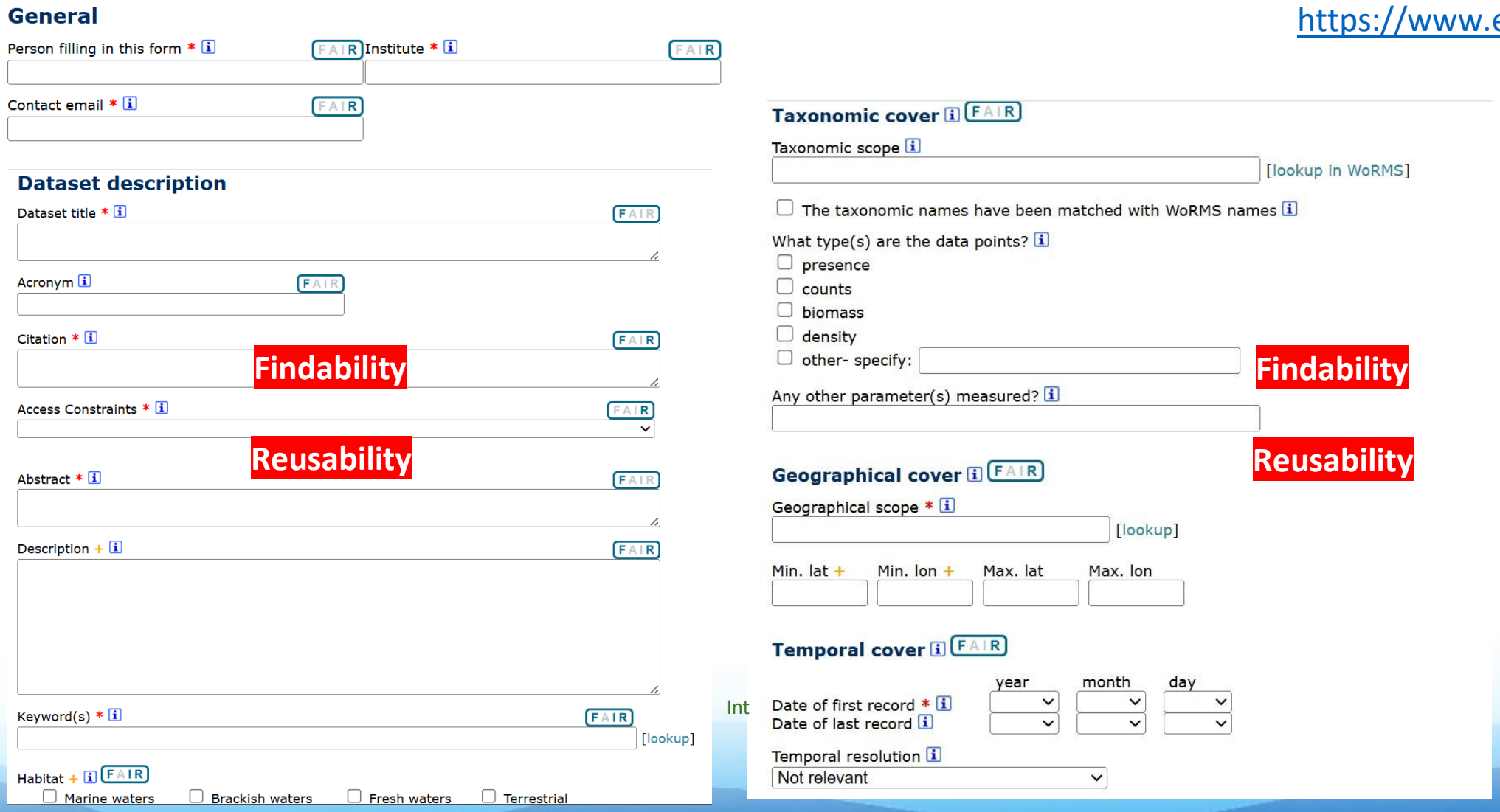
- **Discovery**
  - Title
  - Abstract
  - Keywords
  - Contributors
  - External links
  - Project data
- **Administrative**
  - Contact name
  - Contact affiliation
  - Contact email
  - Contact ID
  - Citation
  - License
- **Descriptive**
  - Taxonomic coverage
  - Geographic coverage
  - Temporal coverage
  - Parameters measured
  - Methodology
  - Instruments, platforms

Integrated Publishing Toolkit (IPT)

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

Example catalogue metadata form

https://www.eurobis.org/submit



Example catalogue metadata form

https://www.eurobis.org/submit

Measurements * FAIR

Parameter | Unit | Method/Protocol

Remov

Add

The dataset FAIR

Indicate how we can obtain or reference your dataset *

Findability

Upload dataset file (max. 10MB)

Choose file No file chosen

Request DOI

Reusability

File format + FAIR

Status of the data collection * FAIR

Basis of the distribution records

Inte

People involved FAIR

Name and contact of person | Institute | Role

Contact
Co-ordinator
Data creator
Data owner

Remov

Add

Reference FAIR

Findability

Reference(s) of publication(s) that was based on this dataset (one reference per textfield) +

Remove

Add

Reusability

Reference(s) of publication(s) that describe this dataset in detail (one reference per textfield) +

Remove

Add

Catalogue metadata

Integrated Marine Information System (IMIS)

Arqueomonitor project: Abundance of the Bucentaure and Fougueux shipwrecks benthic communities, Cadiz Bay (Spain) 2012-2013

Create DOI

Citation

González Duarte, M. M.; Bethencourt Núñez, M.; (2019): Arqueomonitor project: Abundance of the Bucentaure and Fougueux shipwrecks benthic communities, Cadiz Bay (Spain) 2012-2013. https://marineinfo.org/id/dataset/6251

Contact: Bethencourt Núñez, Manuel

Access data

EMODnet EurOBIS OBIS

Archived data

IPT

Availability: This dataset is licensed under a Creative Commons Attribution 4.0 International License.

Special collections:

available through EurOBIS [619]

EMODNET [552]

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

Catalogue metadata

Ecological Metadata Language (EML)

→ the «emk.xml» file in a DwC-A

```
<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/1.2/eml.xsd"
         packageId="https://ipt.vliz.be/eurobis/resource?id=arqueomonitor/v1.0" system="http://gbif.org" scope="system"
xml:lang="en">

    <dataset>
        <alternateIdentifier>https://ipt.vliz.be/eurobis/resource?r=arqueomonitor</alternateIdentifier>
        <title xml:lang="en">Arqueomonitor project: Abundance of the Bucentaure and Fougueux shipwrecks benthic communities,
            <creator>
        <individualName>
            <givenName>Manuel Maria</givenName>
        <surName>González Duarte</surName>
        </individualName>
        <organizationName>Universidad de Cadiz</organizationName>
            </creator>
            <metadataProvider>
        <individualName>
            <givenName>Cristian</givenName>
        <surName>Muñoz Mas</surName>
        </individualName>
        <organizationName>Balearic Islands Coastal Observing and Forecasting System</organizationName>
            </metadataProvider>
            <associatedParty>
        <individualName>
            <givenName>Tomás</givenName>
        <surName>Fernández Montblanc</surName>
        </individualName>
        <organizationName>Universidad de Cadiz</organizationName>
        <role>pointOfContact</role>
            </associatedParty>
            <associatedParty>
        <individualName>
            <givenName>Manuel</givenName>
        <surName>Bethencourt Núñez</surName>
        </individualName>
        <organizationName>Universidad de Cadiz</organizationName>
        <role>pointOfContact</role>
            </associatedParty>
        <pubDate>
            2024-02-27
        </pubDate>
        <language>en</language>
            <abstract>
                    <para>This dataset contains abundance data of the benthic community around the shipwrecks Bucentaure and
            </abstract>
```

DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!