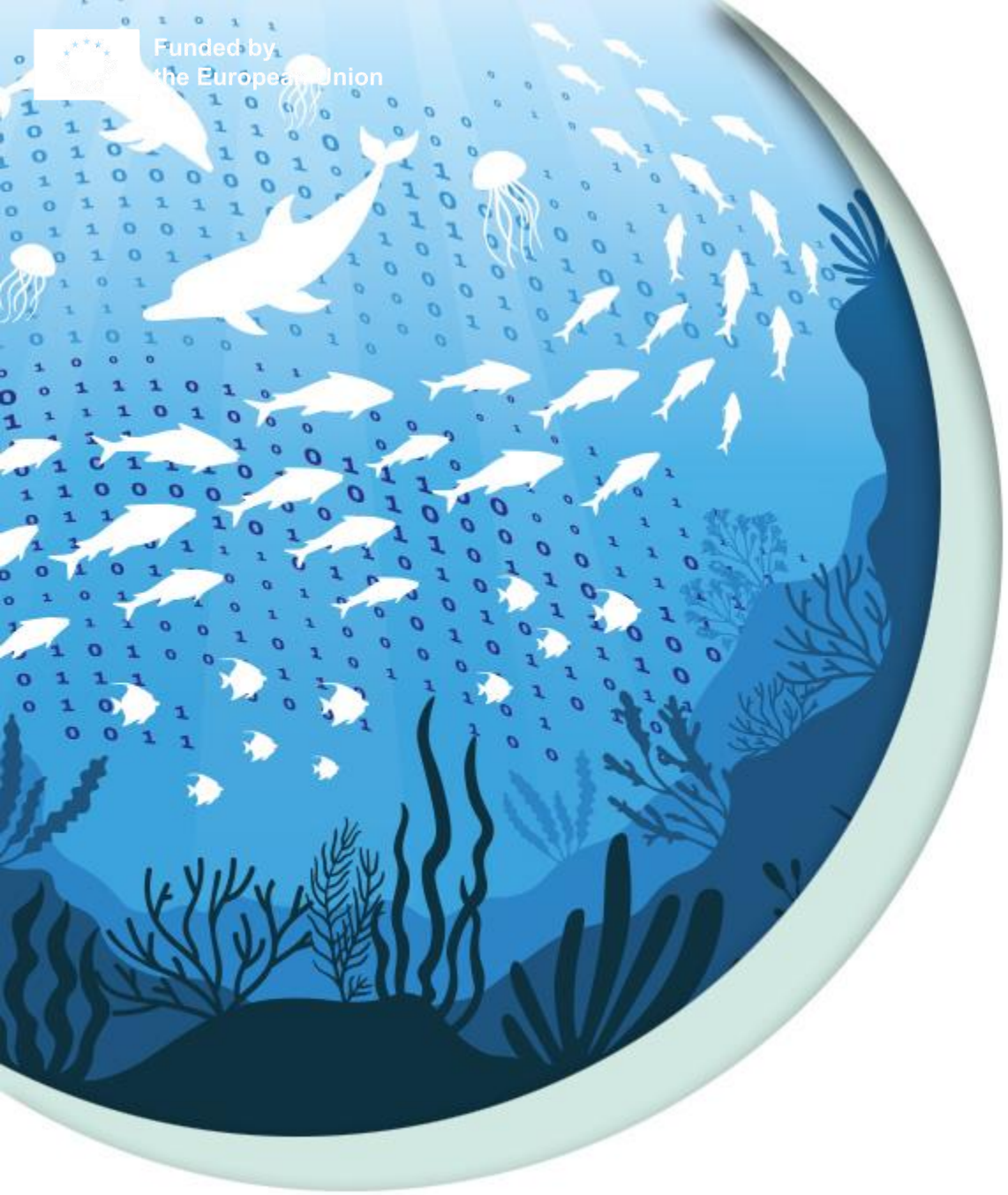


Funded by
the European Union

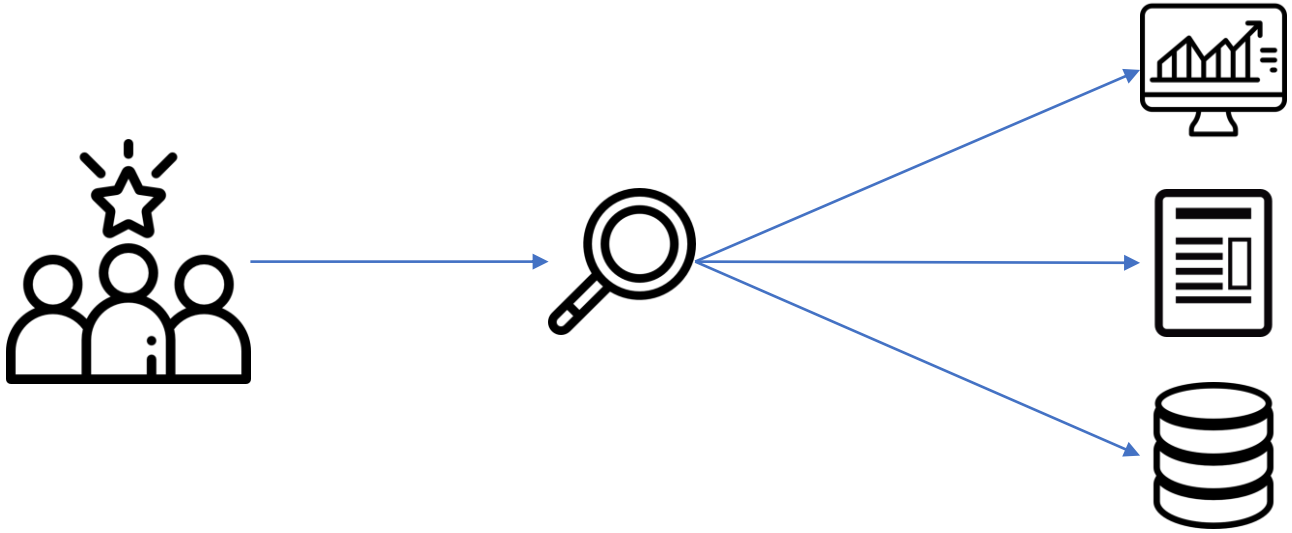
DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

DTO-BioFlow data training workshop:

Darwin Core

State of the art: Data is scattered

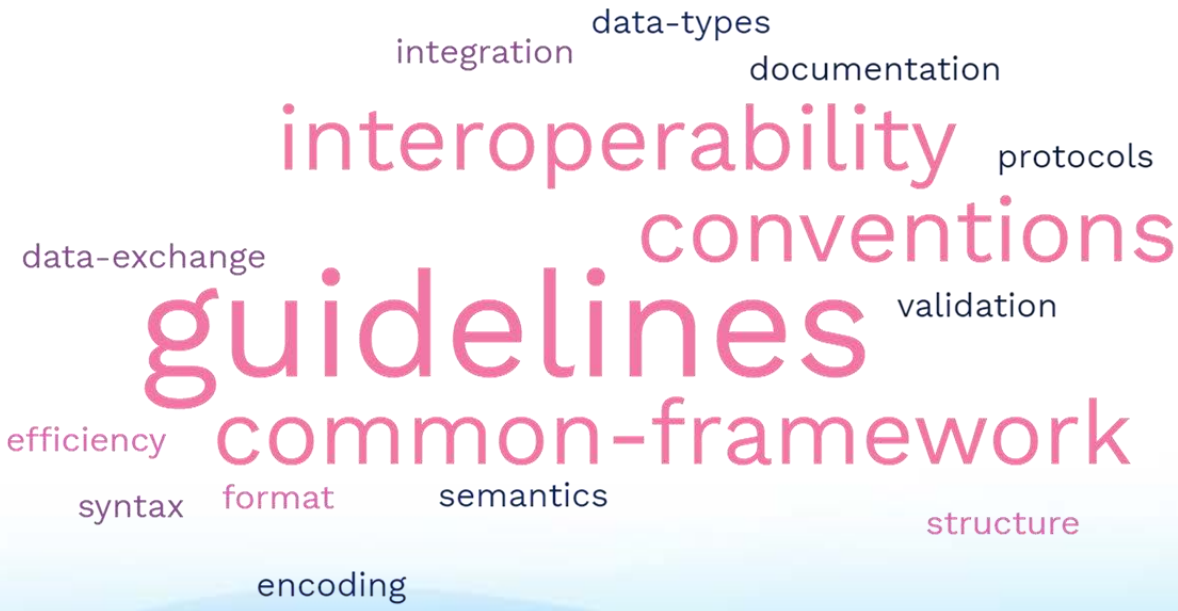


under the current system. Students in PhD programmes spend up to 80% of their time on 'data munging', fixing formatting and minor mistakes to make data suitable for analysis – wasting time and talent. With 400 such students, that would amount to a monetary waste equivalent to the salaries of 200 full-time 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.

FAIR data: Data standards

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



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

FAIR data: Data standards

✗

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

Darwin Core: The scope

Darwin Core is a standard maintained by the Darwin Core Maintenance Interest Group. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to **facilitate the sharing of information about biological diversity** by providing identifiers, labels, and definitions. Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.

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

Darwin Core: The terms

List of terms

- **Standardized**
- **Maintained**

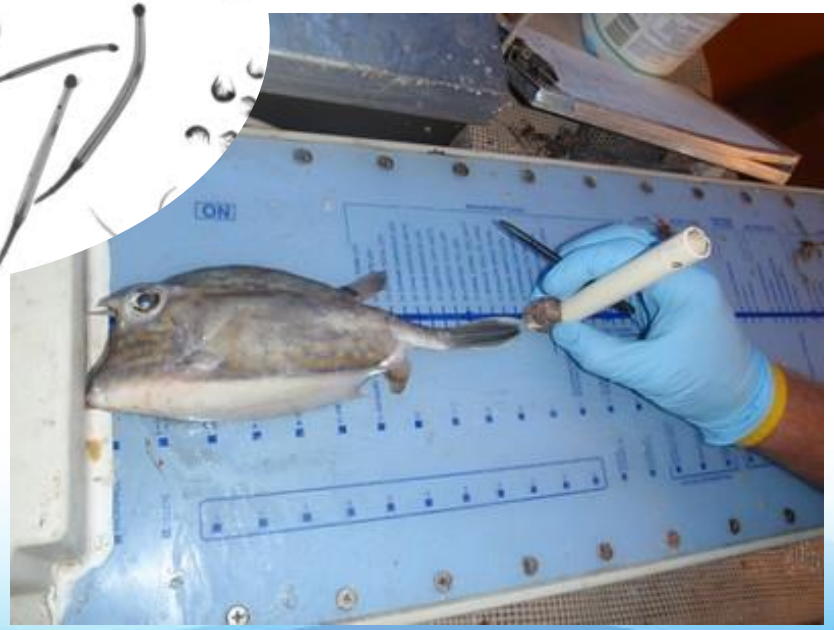
occurrenceID | catalogNumber | recordNumber | recordedBy | recordedByID | individualCount
organismQuantity | organismQuantityType | sex | lifeStage | reproductiveCondition | caste
behavior | vitality | establishmentMeans | degreeOfEstablishment | pathway
georeferenceVerificationStatus | occurrenceStatus | associatedMedia | associatedOccurrences
associatedReferences | associatedTaxa | otherCatalogNumbers | occurrenceRemarks

basisOfRecord	
Identifier	http://rs.tdwg.org/dwc/terms/basisOfRecord
Definition	The specific nature of the data record.
Comments	Recommended best practice is to use a controlled vocabulary such as the set of local names of the identifiers for classes in Darwin Core.
Examples	`MaterialEntity`
	`PreservedSpecimen`
	`FossilSpecimen`

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

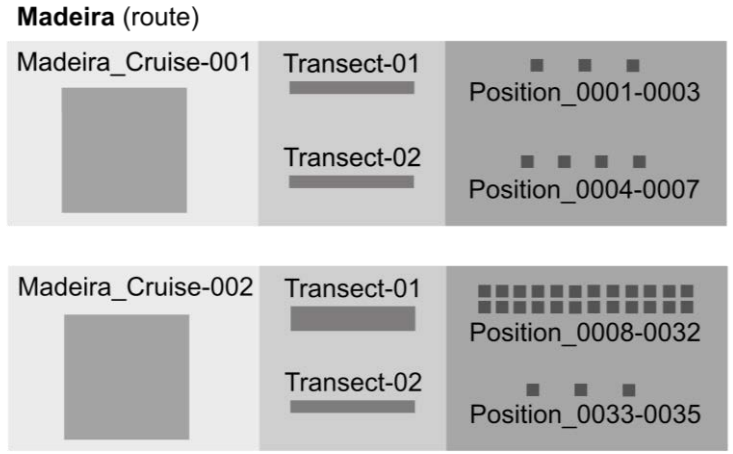
Darwin Core: The concepts

- Event
- Occurrence
- Measurement or Fact



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

Darwin Core: The concepts



Event:

An **action** that occurs at a particular **place** and **time**.

Event	Class
Identifier	http://rs.tdwg.org/dwc/terms/Event
Definition	An action that occurs at some location during some time.
Comments	
Examples	a specimen collecting event
	a camera trap image capture
	a marine trawl

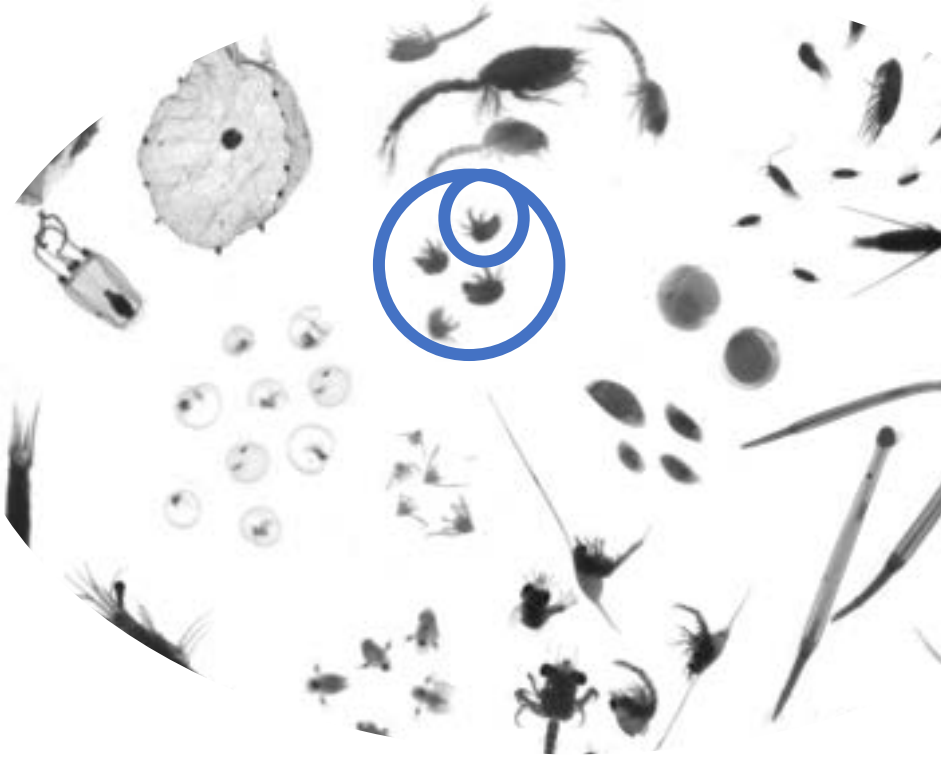
VLIZ

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

Darwin Core: The concepts

Occurrence:

An existence of an **organism** (or **homogeneous** group of organisms) at a particular **place** and **time.**



Occurrence	Class
Identifier	http://rs.tdwg.org/dwc/terms/Occurrence
Definition	An existence of a dwc:Organism at a particular place at a particular time.
Comments	
Examples	a wolf pack on the shore of Kluane Lake in 1988
	a virus in a plant leaf in the New York Botanical Garden at 15:29 on 2014-10-23
	a fungus in Central Park in the summer of 1929

VLIZ

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

Darwin Core: The concepts

https://www.flickr.com/photos/myfwc/12677388

- Measurement or Fact:

  A known **characteristic** of **something.**

MeasurementOrFact	Class
Identifier	http://rs.tdwg.org/dwc/terms/MeasurementOrFact
Definition	A measurement of or fact about an rdfs:Resource (http://www.w3.org/2000/01/rdf-schema#Resource).
Comments	Resources can be thought of as identifiable records or instances of classes and may include, but need not be limited to instances of dwc:Occurrence, dwc:Organism, dwc:MaterialEntity, dwc:Event, dcterms:Location, dwc:GeologicalContext, dwc:Identification, or dwc:Taxon.
Examples	`the weight of a dwc:Organism in grams`
	`the number of placental scars`
	`surface water temperature in Celsius`

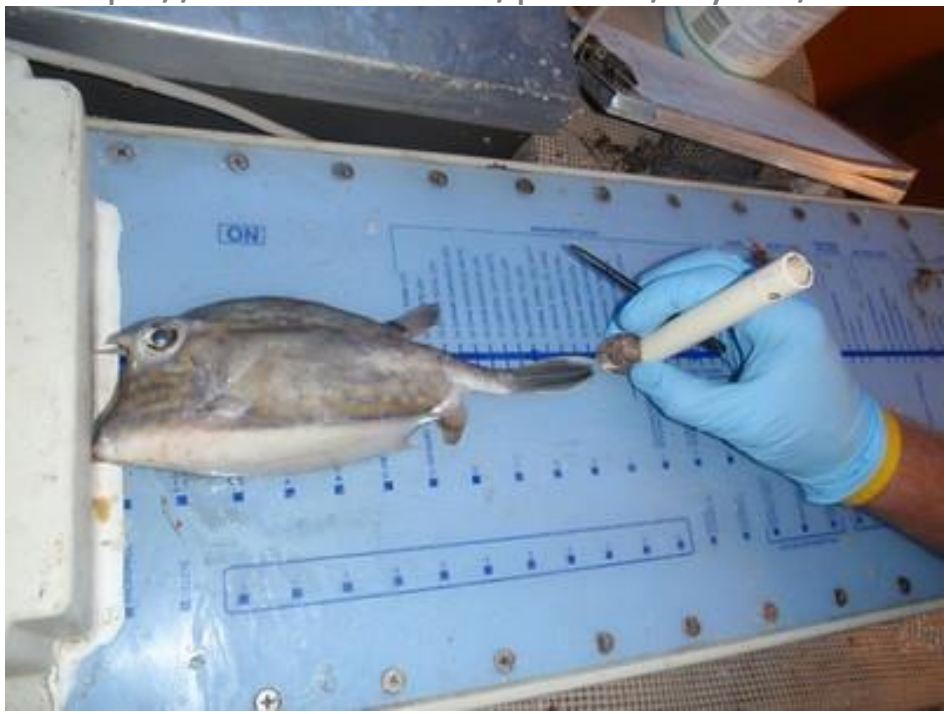
VLIZ

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

Darwin Core: The concepts

- Measurements: Quantitative
  - Water temperature
  - Duration of net tow
  - Length of the organism
- Facts: Qualitative
  - Habitat type
  - Sampling instrument name
  - Life stage of the organism

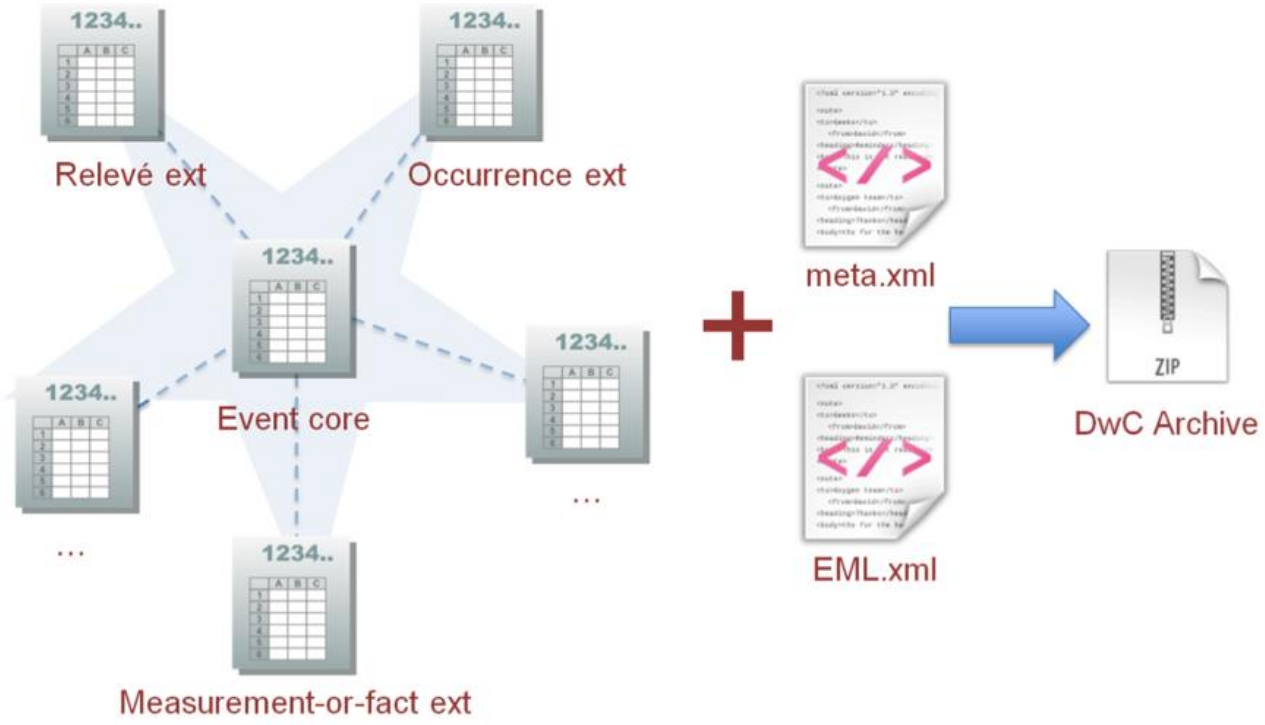
https://www.flickr.com/photos/myfwc/12677388



VLIZ

Darwin Core Archives: The data package

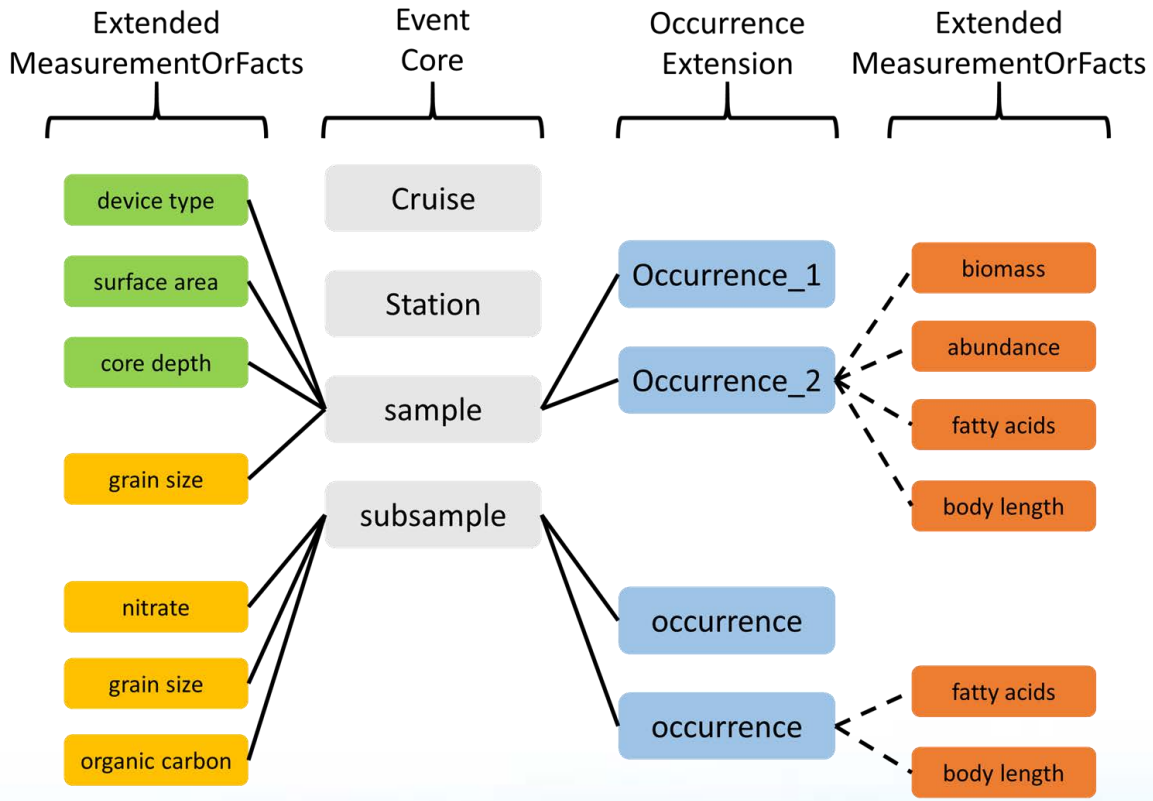
- Data tables
  - Core
  - Extensions
- Meta.xml
- Eml.xml



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

Darwin Core Archives: The data package

- **Data tables**
  - **Core**
  - **Extensions**
- Meta.xml
- Eml.xml

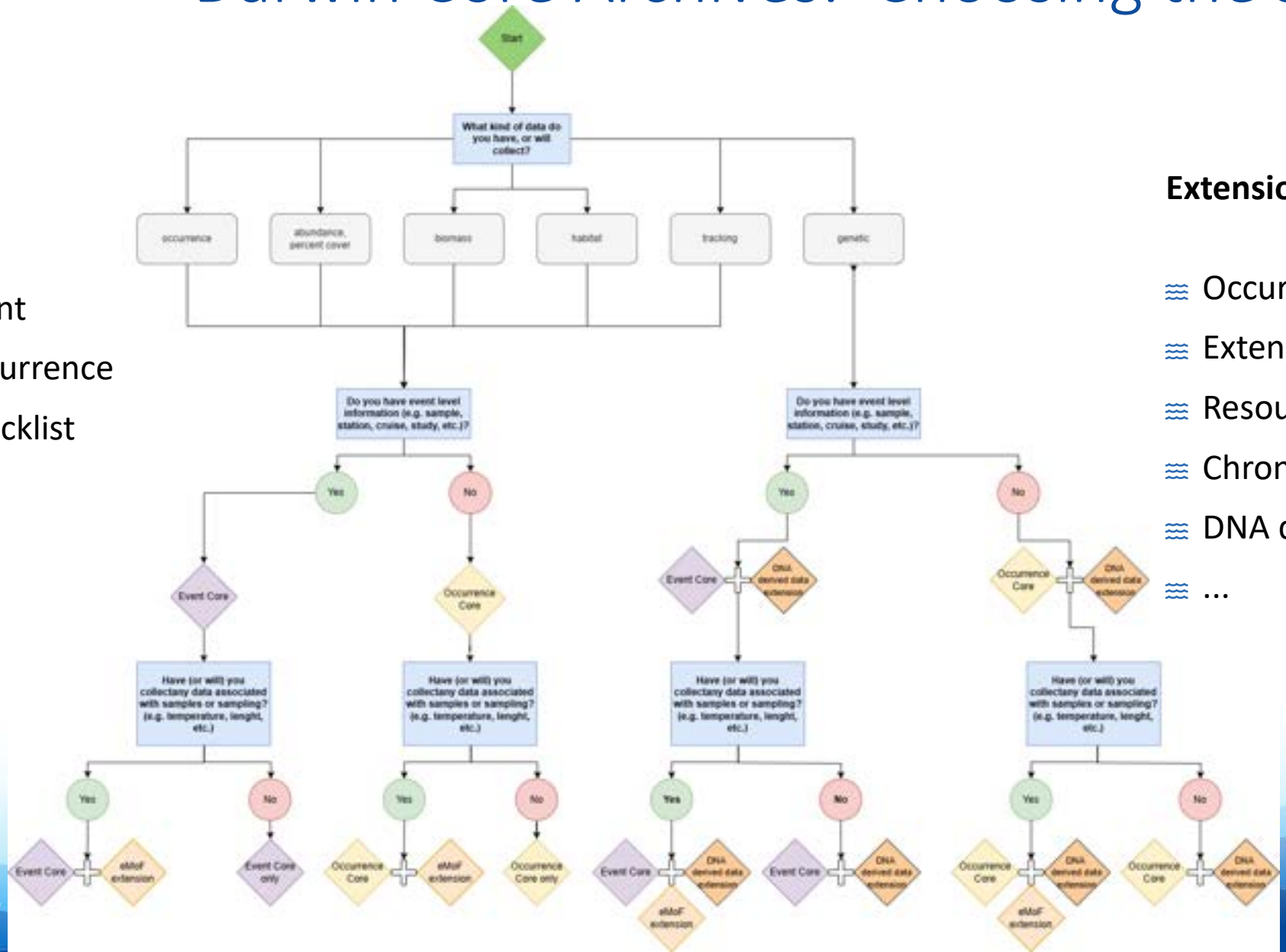


DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Darwin Core Archives: Choosing the structure

Cores

- Event
- Occurrence
- Checklist

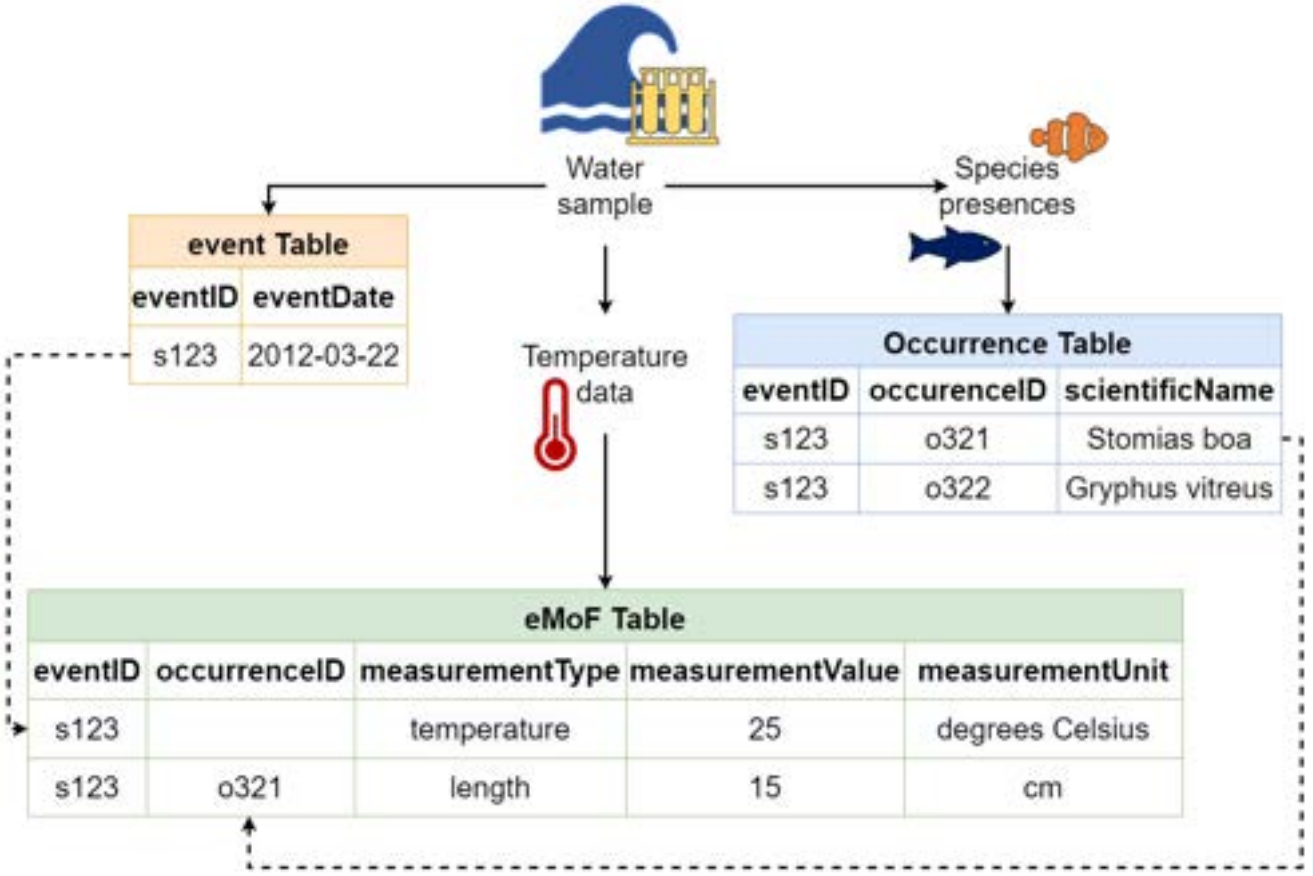


Extensions

- Occurrence
- Extended Measurement Or Facts
- Resource Relationships
- Chronometric Age
- DNA derived data
- ...

Darwin Core Archives: Example

- **Data tables**
  - **Core**
  - **Extensions**
- Meta.xml
- Eml.xml



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

Darwin Core Archives: Example

- Data tables
  - Core
  - Extensions
- **Meta.xml**
- Eml.xml

```
<archive xmlns="http://rs.tdwg.org/dwc/text/" metadata="eml.xml">
  <core encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n" fieldsEnclosedBy="">
    <files>
      <location>occurrence.txt</location>
    </files>
    <coreid index="0" />
    <field index="1" term="http://rs.tdwg.org/dwc/terms/basisOfRecord"/>
    <field index="2" term="http://rs.tdwg.org/dwc/terms/occurrenceID"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/occurrenceRemarks"/>
    <field index="4" term="http://rs.tdwg.org/dwc/terms/recordedBy"/>
    <field index="5" term="http://rs.tdwg.org/dwc/terms/behavior"/>
    <field index="6" term="http://rs.tdwg.org/dwc/terms/occurrenceStatus"/>
    <field index="7" term="http://rs.tdwg.org/dwc/terms/previousIdentifications"/>
    <field index="8" term="http://rs.tdwg.org/dwc/terms/eventID"/>
    <field index="9" term="http://rs.tdwg.org/dwc/terms/identifiedBy"/>
    <field index="10" term="http://rs.tdwg.org/dwc/terms/identificationQualifier"/>
    <field index="11" term="http://rs.tdwg.org/dwc/terms/scientificNameID"/>
    <field index="12" term="http://rs.tdwg.org/dwc/terms/scientificName"/>
  </core>
  <extension encoding="UTF-8" fieldsTerminatedBy="\t" linesTerminatedBy="\n" fieldsEnclosedI
    <files>
      <location>extendedmeasurementorfact.txt</location>
    </files>
    <coreid index="0" />
    <field index="1" term="http://rs.tdwg.org/dwc/terms/measurementID"/>
    <field index="2" term="http://rs.tdwg.org/dwc/terms/occurrenceID"/>
    <field index="3" term="http://rs.tdwg.org/dwc/terms/measurementType"/>
    <field index="4" term="http://rs.iobis.org/obis/terms/measurementTypeID"/>
    <field index="5" term="http://rs.tdwg.org/dwc/terms/measurementValue"/>
    <field index="6" term="http://rs.iobis.org/obis/terms/measurementValueID"/>
    <field index="7" term="http://rs.tdwg.org/dwc/terms/measurementUnit"/>
    <field index="8" term="http://rs.iobis.org/obis/terms/measurementUnitID"/>
  </extension>
</archive>
```

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

Darwin Core Archives: Example

- Data tables
  - Core
  - Extensions
- Meta.xml
- **Eml.xml**

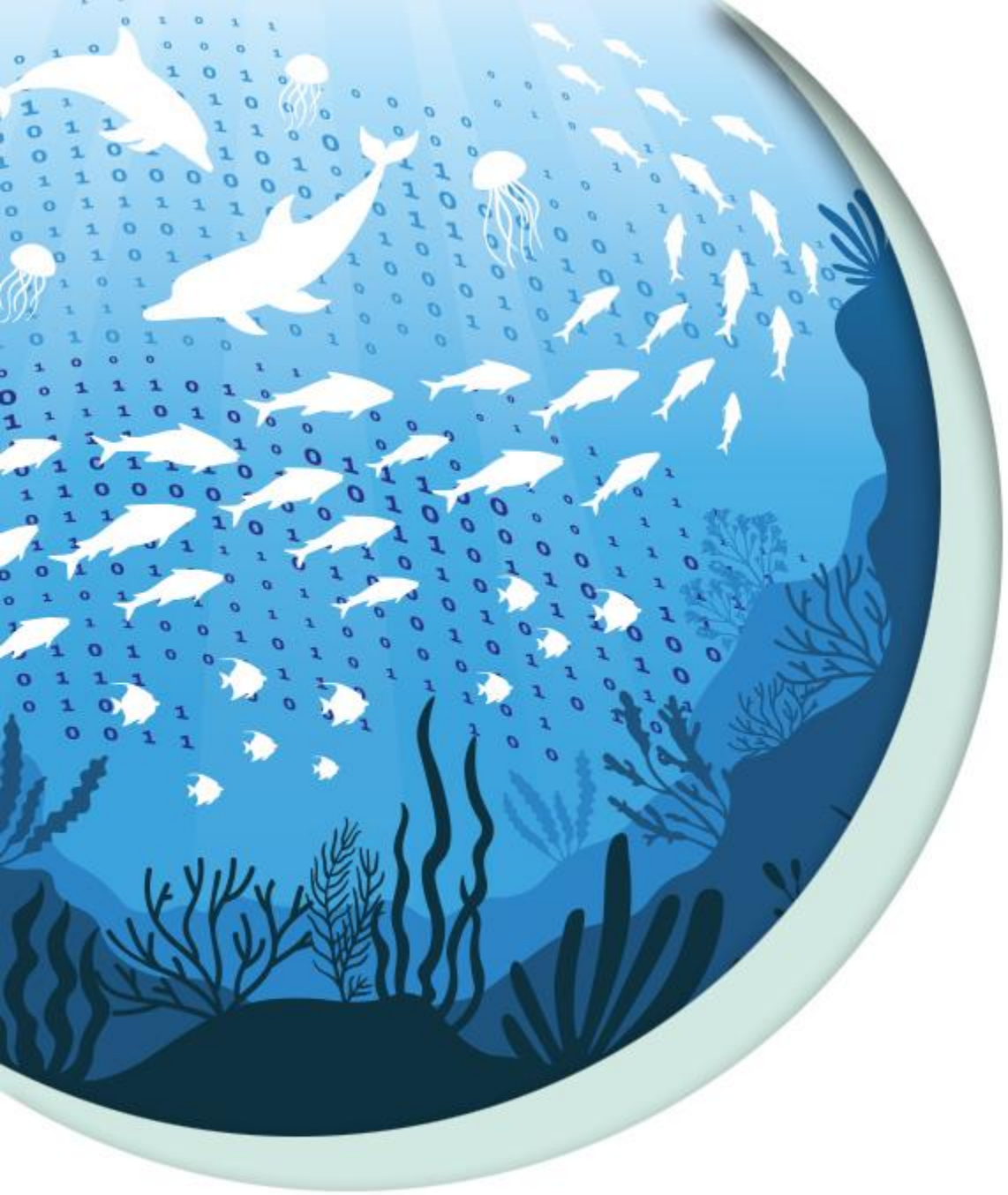
```
<eml:eml xmlns:eml="eml://ecoinformatics.org/eml-2.1.1"
         xmlns:dc="http://purl.org/dc/terms/"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="eml://ecoinformatics.org/eml-2.1.1 http://rs.gbif.org/schema/eml-gbif-profile/1.1
         packageId="cd4f3685-c3dd-4461-894c-b4e94c17585f/v1.3" system="http://gbif.org" scope="system"
         xml:lang="eng">

<dataset>
  <alternateIdentifier>10.17031/tpgjrv</alternateIdentifier>
  <alternateIdentifier>cd4f3685-c3dd-4461-894c-b4e94c17585f</alternateIdentifier>
  <alternateIdentifier>https://www.dassh.ac.uk/ipt/resource?r=dassh-113</alternateIdentifier>
  <title xml:lang="eng">1999-2001 University Marine Biological Station Millport (UMBSM) Clyde Sea Rapid Method
    <creator>
    <organizationName>University Marine Biological Station Millport (UMBSM)</organizationName>
    </creator>
    <metadataProvider>
    <organizationName>The archive for marine species and habitats data (DASSH)</organizationName>
    <address>
      <city>Plymouth</city>
      <country>GB</country>
    </address>
    <electronicMailAddress>dassh.enquiries@mba.ac.uk</electronicMailAddress>
    </metadataProvider>
    <associatedParty>
    <individualName>
      <givenName>Data</givenName>
    <surName>Team</surName>
    </individualName>
    <electronicMailAddress>data@mba.ac.uk</electronicMailAddress>
    <role>user</role>
    </associatedParty>
```

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

Relevant sources

- GBIF Repository of Schemas
- Darwin Core terms
- Darwin Core archive IPT guide
- OBIS manual
- Ecological Metadata Language (EML)
- 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



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!