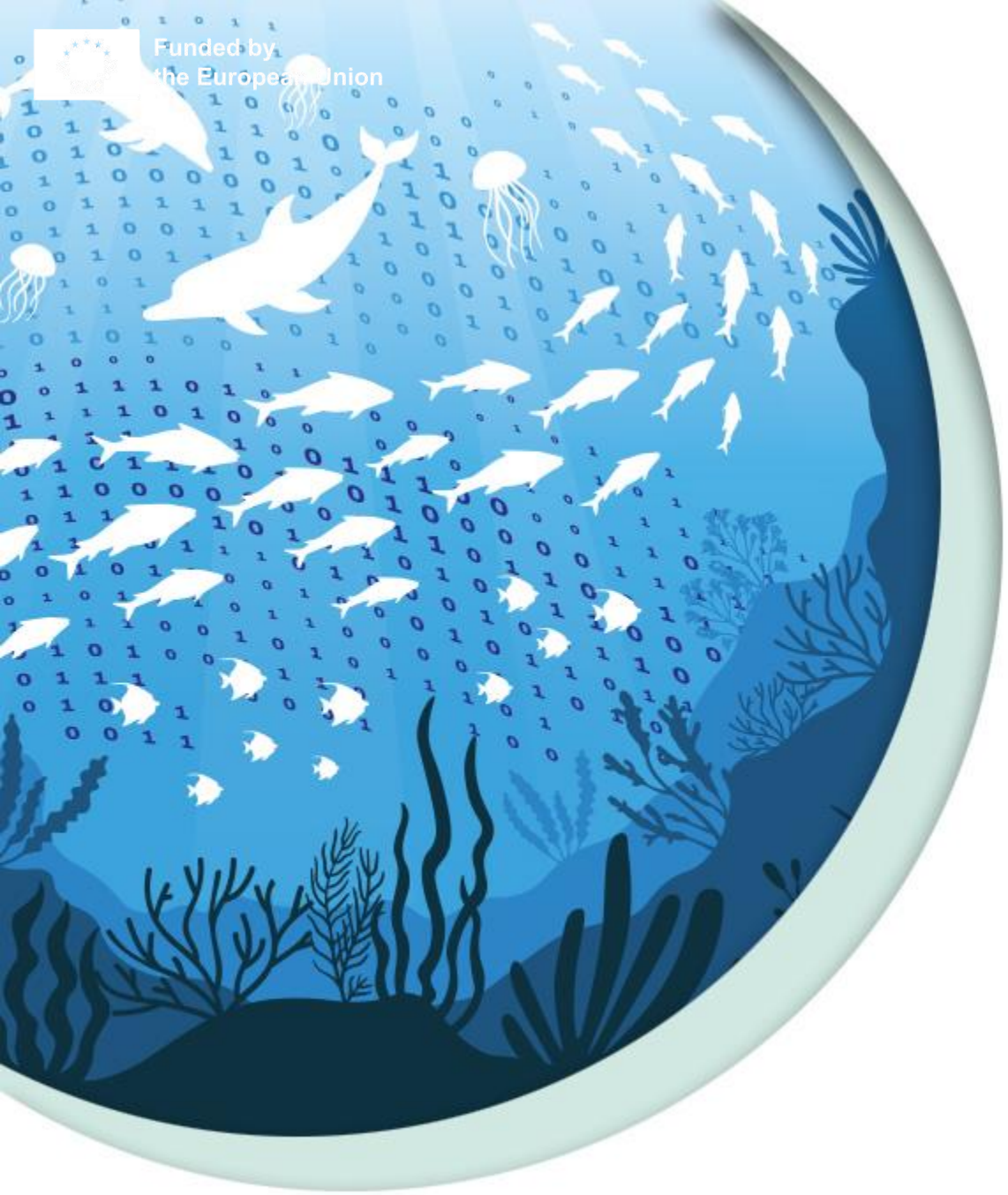


Funded by
the European Union

DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

DTO-BioFlow data training workshop:

Good practices for data capture in spreadsheets

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

SPREADSHEETS

- Most frequently used tool for data collection
- Often used for data storage
- Starting point for the transformation to DwC-A

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains calculated values and plots

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

Common problems with spreadsheets

- **Data structure often optimised for human readability rather than machine readability**
- Data is organised in many different ways
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

Avoid pretty spreadsheets

- Avoid:
  - Using blank rows or columns to visually separate things
  - Using merged/overarching cells
  - Using multiple rows for header names
  - Using colour to convey meaning
  - Using blank cells to indicate repetition of a value
  - Adding multiple tables on the same sheet

	A	B	C	D	E	F	G	H
1		**Location 1**		**Location 2**				
2	**Date**	**Species**	**Count**	**Species**	**Count**		**Location 1**	Oostende Bank
3								51.29785,2.788396
4	02/08/2023	Species A	1	Species A	7			Sandbank
5		Species B	50	Species B	0		**Location 2**	Middelkerke Bank
6		Species C	0	Species C	2			51.28682,2.722723
7								Sandbank
8	03/08/2023	Species A	2	Species A	5			
9		Species B	4	Species B	1			
10		Species C	3	Species C	4			

→ Structure should always be a simple rectangle
+ all data should be captured as text (not just colour)

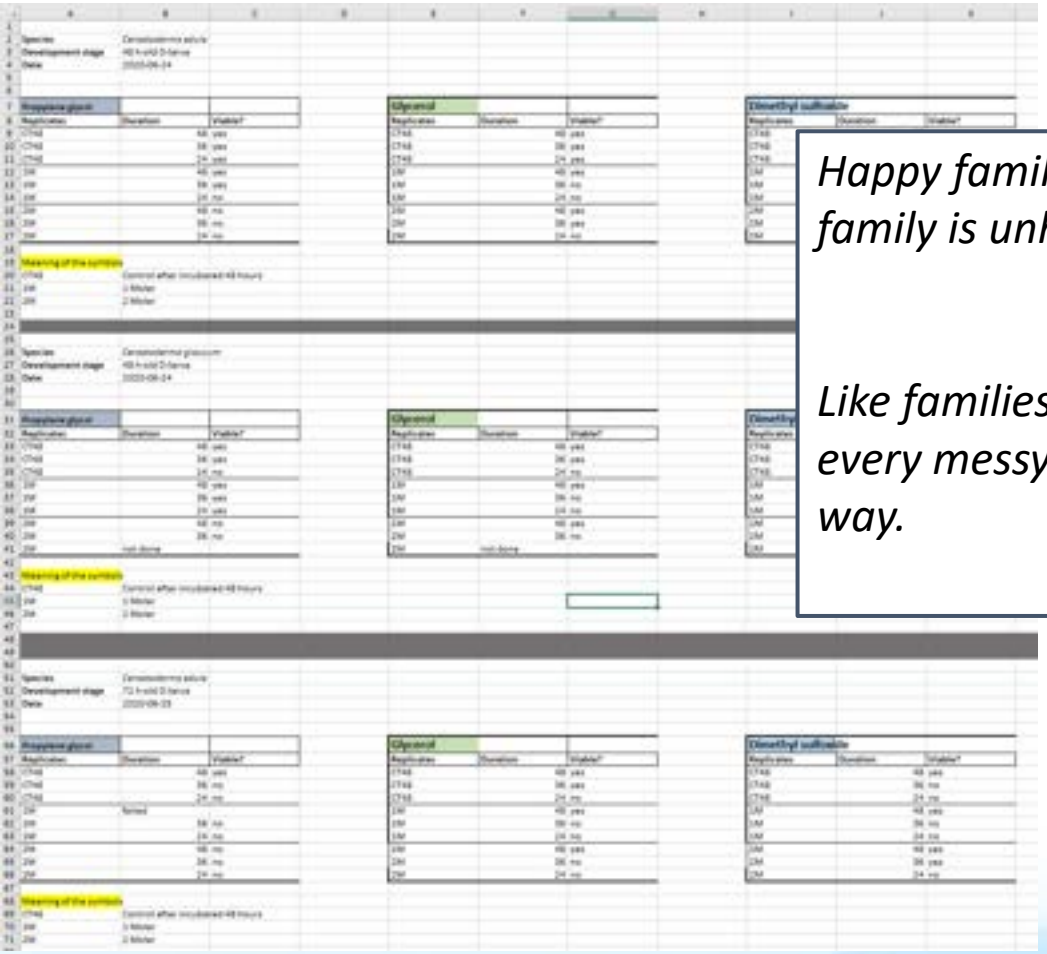
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- **Data is organised in many different ways**
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

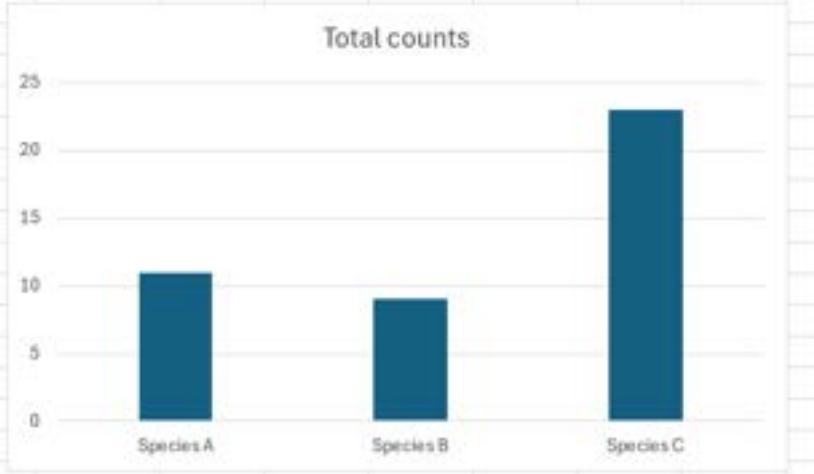
DTO-BioFlow
Integration of biodiversity monitoring data into the Digital Twin Ocean

Tidy data



	A	B	C	D
1		Locations	Species	Count
2	28/03/2024	Site 1	Species A	5
3			Species B	1
4			Species C	7
5		Site 2	Species A	0
6			Species B	4
7			Species C	5
8	29/03/2024	Site 1	Species A	4
9			Species B	3
10			Species C	8
11		Site 2	Species A	2
12			Species B	1

Total counts	Species A	11
	Species B	9
	Species C	23



Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Like families, tidy datasets are all alike but every messy dataset is messy in its own way.

Hadley Wickham

	A	B	C	D	E
1		Location 1		Location 2	
2	Date	Species	Count	Species	Count
3					
4	2023-08-02	Species A	1	Species A	7
5		Species B	50	Species B	0
6		Species C	0	Species C	2
7					
8	2023-08-03	Species A	2	Species A	5
9		Species B	4	Species B	1
10		Species C	3	Species C	4
11					

Location 1	Oostende Bank
	51.29785,2.788396
	Sandbank
Location 2	Middelkerke Bank
	51.28682,2.722723
	Sandbank

Tidy data

- Standard way to structure data
- Three basic principles:
  - each variable is a column
  - each observation is a row
  - each value is a cell

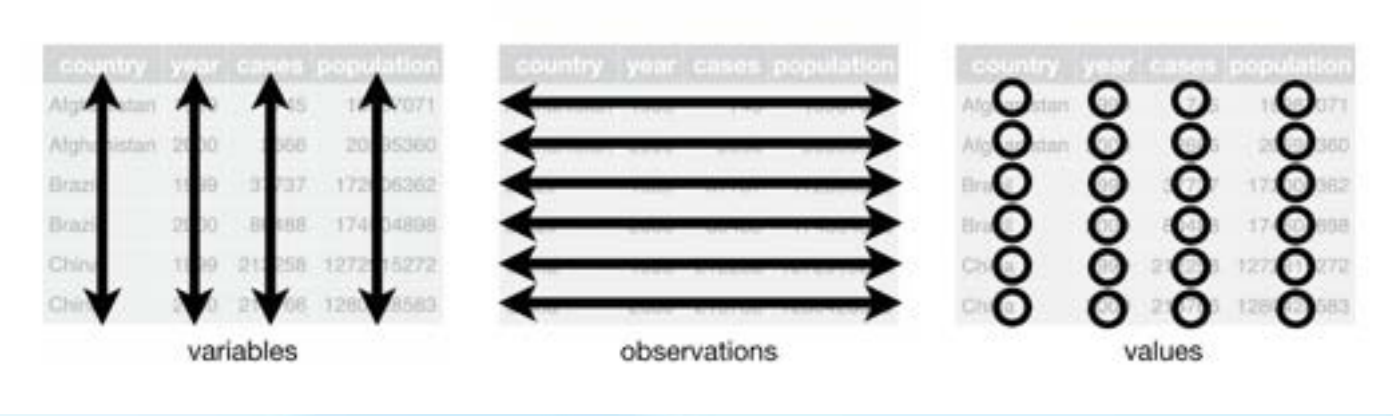


Image from “R for Data Science (2e)”.

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

Tidy data

- Standard way to structure data
- Three basic principles:
  - each variable is a column
  - each observation is a row
  - each value is a cell

Species	Male	Female
Species A	2	2
Species B	1	0
Species C	4	6

Species	Sex	Count
Species A	Male	2
Species B	Male	1
Species C	Male	4
Species A	Female	2
Species B	Female	0
Species C	Female	6

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

Tidy data

- Advantages
  - Increase interoperability
  - Flexible
  - Easy to analyse and visualise
  - Excellent tools (e.g. Tidyverse in R)
  - Add meaning to the structure
  - Most DwC tables are tidy

→ **easier transformation**

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- **Bad variable names**
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

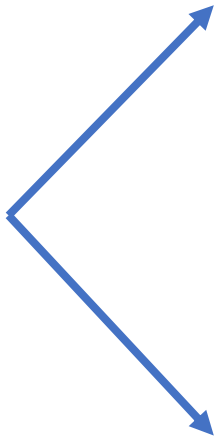
Clear variables

- Good names
  - Descriptive
  - Concise
  - No spaces or special characters

Clear variables

- Good names
- Units
  - In separate column
  - In variable name
  - In data dictionary

Biomass
50 mg/m^3
20 mg/m^3
100 mg/m^3



Dry_weight_biomass_mg_per_m3
50
20
100

Dry_weight_biomass	Dry_weight_biomass_unit
50	mg/m^3
20	mg/m^3
100	mg/m^3

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

Clear variables

- Data dictionary: separate file / sheet to document additional info

	A	B	C	D	E	F	G
1	Start Date	Additive	Conce	Volume	Biological material	Mass of material	
2	2022-01-01	NaCl	50	1	brown seaweed	100	
3	2022-01-01	NaCl	75	1	brown seaweed	200	
4	2022-01-01	NaCl	50	1	red seaweed	100	
5	2022-01-01	Cl	50	2	brown seaweed	100	
6	2022-01-01	Cl	75	2	brown seaweed	200	
7	2022-01-01	Cl	50	2	red seaweed	100	
8							

	A	B	C	D	E
1	Parameter	ParameterType	Description	VocabularyID	Unit
2	Start date	column header	When the material was added to the water+preservative	NA	NA
3	Preservative	column	Type of preservative used	NA	NA
4	Concentration	column	Concentration of the preservative	http://vocab.nerc.a	percent
5	Volume	column	Total liquid volume	NA	litre
6	Biological material	column	Name of biological material	NA	NA
7	Mass of material	column	Mass of biological material	NA	gram
8	brown seaweed	cell value	Fucus distichus Linnaeus, 1767	https://www.marin	NA
9	red seaweed	cell value	Furcellaria lumbricalis (Hudson) J.V.Lamouroux, 1813	https://www.marin	NA
10	NaCl	cell value	sodium chloride	http://vocab.nerc.a	NA
11	Cl	cell value	chlorine	http://vocab.nerc.a	NA

Data dictionary
(example from "FAIR data for marine biologists' OceanTraining course)

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

Clear variables

- Good names
- Units
  - In separate column
  - In variable name
  - In data dictionary
- Data dictionary: separate file / sheet to document additional info
- Consistent names across files

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Bad variable names
- **Values are not consistent**
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- Spreadsheet contains also calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

Consistent values

- All values in a single column should be in the same format/ written in the same way/ abbreviated in the same way
  - E.g. “juvenile”, “J”, “juv.”, “juv”
  - E.g. “2 Oct-23”, “2023-10-02”, “2/10/2023”
  - → Pick one format and stick to it
- Do not mix data types
  - E.g. writing “between 5 and 10” in a column with numerical values

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

Consistent values

- Missing values
  - Different recommendations
    - Empty cells, NULL, NA
    - Chose 1, stick to it, and document it in your data dictionary/readme file
  - 0 != missing value

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Variable names are not clear
- Bad variable names
- **Contradicting information caused by redundancy**
- Spreadsheet software messed up the data
- Spreadsheet contains also calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

Normalisation

- Normalised → Each type of observational unit is a table
- To normalise or not? → Depends on the purpose
- Normalised data preferred for data storage
  - Avoids reduncancy
    - Efficient data storage
    - Reduces chances of errors
  - Common column needed to link
    - Preferably an ID (which does not hold information in itself)

Employees' Skills

Employee ID	Employee Address	Skill
426	87 Sycamore Grove	Typing
426	87 Sycamore Grove	Shorthand
519	94 Chestnut Street	Public Speaking
519	96 Walnut Avenue	Carpentry

Example of an update anomaly in an unnormalized table

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- **Spreadsheet software messed up the data**
- Spreadsheet contains also calculated values and plots
- → Researchers / data managers spend a lot of time data wrangling

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

Spreadsheet software common issues

- Beware of:
  - Auto-formatting
  - Auto-fill
  - Number separators
  - Encoding issues
  - Maximum size
  - Loss of zeros
  - Overwriting

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

Common problems with spreadsheets

- Data structure often optimised for human readability rather than machine readability
- Data is organised in many different ways
- Bad variable names
- Values are not consistent
- Contradicting information caused by redundancy
- Spreadsheet software messed up the data
- **Spreadsheet contains calculated values and plots**
- → Researchers / data managers spend a lot of time data wrangling

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

Safeguarding the data file

- Always use a copy of the data file for analysis/ visualisation/ transformation
- Write-protect
- Back-up
- Save in a plain-text format

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

A good spreadsheet

- Data structure is optimised for machine readability
- Data is organised in a tidy structure
- Good variable names, clear meanings, units documented
- Values are consistent
- No redundancy (or no contrasting information)
- Basic quality inspection done
- Spreadsheet contains only the raw data and is saved as CSV/TSV
- → Easy transformation to standard format
- → Researchers spend a lot of time ~~data wrangling~~ doing science

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

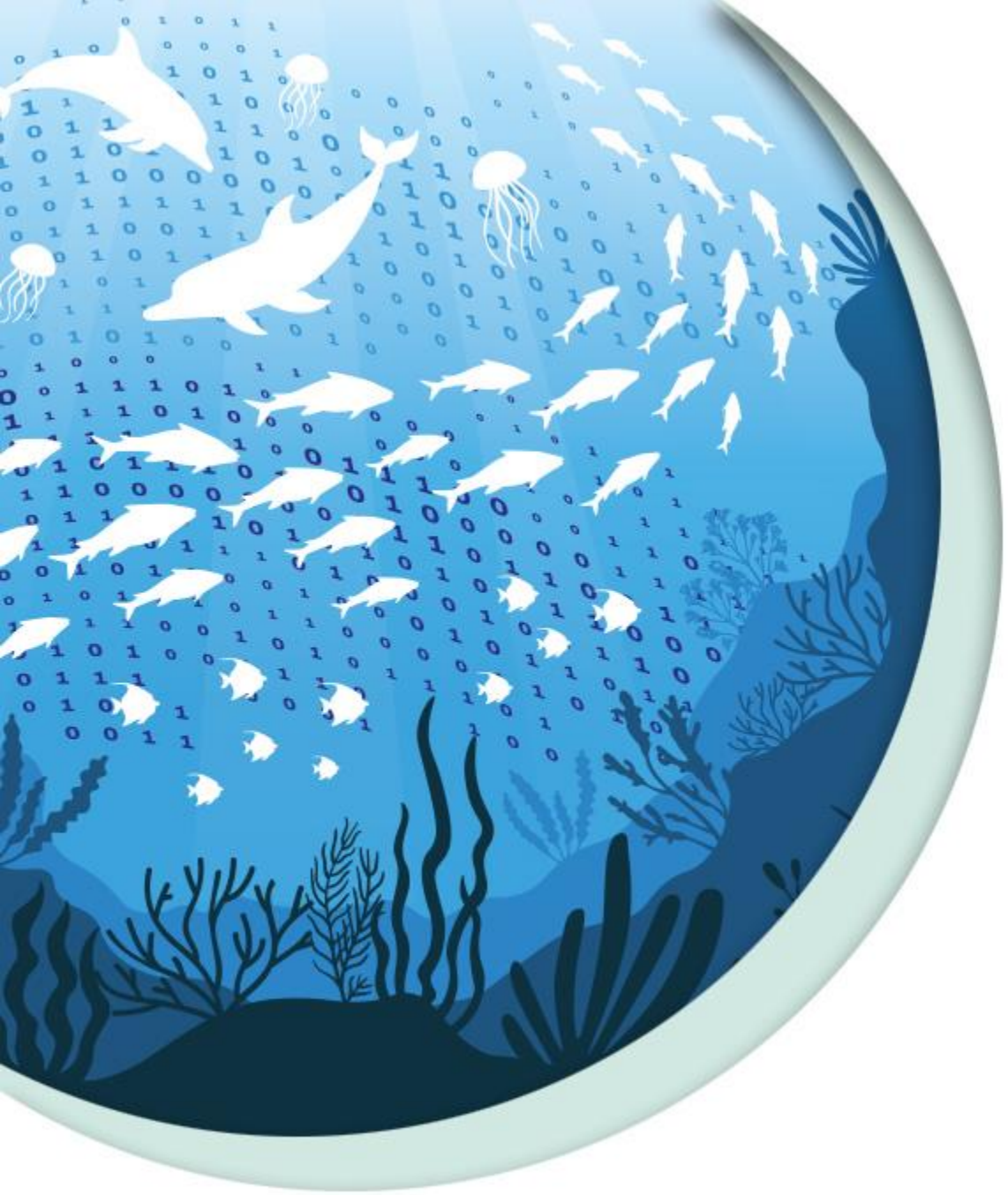
A good spreadsheet

- Ideally, these good practices are applied from the start (data entry)
- Alternatively, we will need to transform 'messy' data
  - Always work on a copy
  - Keep track of the transformation steps
    - E.g. README file
  - Preferably, use a tool that automatically tracks the transformation steps and allows the transformation to be easily repeated (e.g. script, OpenRefine)

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

Resources

- OTGA course Biological Data Management: https://classroom.oceanteacher.org/course/view.php?id=918 (Module 3. Best practices for data capture)
- Wickham, H. . (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10
- Broman, K. W. and Woo, K. H. (2018) 'Data Organization in Spreadsheets', *The American Statistician*, 72(1), pp. 2–10. https://doi.org/10.1080/00031305.2017.1375989
- R for Data Science: Data tidying: https://r4ds.hadley.nz/data-tidy.html
- Python4DS: Tidy data: https://aeturrell.github.io/python4DS/data-tidy.html
- Data Cleaning with OpenRefine for Ecologists: https://datacarpentry.org/OpenRefine-ecology-lesson/index.html



DTO-BioFlow

Integration of biodiversity monitoring data into the Digital Twin Ocean

THANKS!