**Assignment DataFest RWS biodiversity data**

**Introduction**

Rijkswaterstaat monitors all kinds of biological variables, in all its waterways, every year. Likewise in Zeeland. You are going to look into several of the corresponding datasets:

* Macrozoobenthos
* Phytoplankton
* Diatoms
* Macroalgae (optional)
* Aquatic Plants (optional)

These datasets consist of measurements from several locations, over several years. The goal is to gain insight into these datasets and thus see whether climate change has an effect on the biology in the Zeeland Delta. This can, among others be done as follows:

1. Create an (interactive) map with a graph for each location to show the course of time in terms of species composition
2. Map disappeared/disappearing and emerging taxa
3. Elaborate on the results of this in more detail at your own discretion

The context and the assignment is explained in more detail in this document.

**Dataset preprocessing**

The datasets are validated and ready to use, but some preprocessing is definitely needed to better understand, and use the data.

Update the datasets

1. View the metadata with the explanation of the columns. The following information is relevant, in any case:
   1. Location data (name and coordinates)
   2. Collection date
   3. Collection\_Reference: each sample has a unique number. This also includes information about the sampling such as the sampling method. In the Measurement package\_Code column you can find whether it concerns taxa data or metadata.
   4. The taxon name is given in the Parameter\_Specification column.
   5. Because only part of the sample is often counted per sample, the Value\_Calculated column indicates how many of each taxon were in the sample. The unit is also given.
   6. The limit symbol “>” indicates that a taxon was found, but not counted. Think about what you want to do with this!
2. Additional information for all taxa can be found in the TWN list (Taxa Waterbeheer Nederland). For more information see: <https://acceptatie.aquo.nl/index.php/Vraag_%26_Antwoord_TWN>. Link the TWN list to the dataset.
   1. There have been changes in naming over the years. The TWN list indicates the most recent name with status code 10. If a name is out of date, the status code has been changed to 20 and under refername you can see what the new name is. Update all taxa in the dataset to the most recent name.
   2. A taxon can be identified at different levels: species-genus-familia-ordo-classis-phylum. In the TWN list you can always find the taxon level (column taxonlevel) of the taxon and what the parent taxon is (column parentname). To compare the data, look at taxa of the same level. Species data provides the most information, but is less easy to compare due to the large number of different species. In addition, not all taxa have been identified at this level. The associated taxon group is also given in the TWN list. By looking at the data at this level, a more global overview of the taxon composition can be obtained. For example, for macrozoobenthos it is useful to look at species, genus and taxon group level. Add columns with the genus and taxongroup to the dataset.

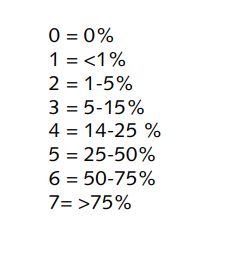
Cleaning and other data wrangling

Obviously, once you updated the datasets according to the TWN list, additional activities may be needed, for instance dealing with missing values, feature engineering and transformations. That depends on what you encounter during data understanding and what is needed in relation to the goals you set for modeling the data.

Dealing with the target values

1d., under “Update the datasets”, mentions counting the taxa. How taxa are counted varies per dataset. It is best to work with value\_calculated. Value\_measured are the counts in the sample itself, but because, for example, only part of a very large sample was counted, this has been corrected at value\_calculated. So value\_calculated gives the result for the whole sample:

1. Macrozoobenthos: number in the sample. However, there are different sampling methods (see classification\_code) with an area (quantity\_code and value\_calculated). For comparability between the methods, it is good to convert everything into n/m2 with the surface area, for example.
2. Diatoms: % as a unit. This means the percentage of the total found in the sample. For example, if the sample consists of 200 dishes and 100 dishes consist of one species, then that species scores 50%.
3. Phytoplankton\_diatoms: Use value\_calculated with the unit\_calculated (n/ml or n/l)
4. Macroalgae: This is the coverage percentages of the area sampled. This is given in percentages, but occasionally also with a score from 0 to 7. This score also represents a percentage of coverage, i.e.



1. Aquatic plants: is given in coverage (see Size\_code). See Macroalgae.

**Assignment**

1. Create an (interactive) map with a graph for each location to show the course of time in terms of species composition. The data consists of many years and sample locations. To gain more insight into this, various types of graphs can be made. Examples are:
   1. By location or by body of water (these locations start with the same abbreviation but end with a different number)
   2. Display multiple years in one chart
   3. Present taxa at different levels: species, genus or taxon group.
   4. Distinction in sampling method. Different sampling methods yield different taxa.
2. Map disappeared/disappearing and emerging taxa. You can also look into specific species. For example, it can be interesting to show species that have come to the area over the years in separate graphs. The same can be done for species that have disappeared.
3. Elaborate on the results of this in more detail at your own discretion. There are a lot of possibilities for modeling for instance
   1. Try and find relations between numbers of species, or numbers of specific species, with water quality data (obviously, here several sub problems need to be solved such as data processing for water quality over time, at specific locations)
   2. Try and find relations for numbers of species, or numbers of specific species between different locations

Most of all, if you decide to focus on modeling, we encourage you to come up you’re your own ideas on what, and how to model!