Derivation of Typical Assemblages

Jonathan Jupke

Oktober 02, 2020

Contents

1 Introduction

Here we describe the methods we used to derive typical assemblages of macro-invertebrates and diatoms for selected river types across Europe. We also describe and briefly discuss the results.

2 Harmonizing taxa names

International diatom occurrence data sets require extensive harmonization because of the taxonomic resolution differing between data sets, different working groups using different nomenclatures, identification errors, and ongoing changes to the accepted nomenclature (Kahlert *et al.* 2020). Harmonization can reduce overall taxonomic resolution but also improve the detection of large-scale spatio-temporal patterns (Lee *et al.* 2019). We compared all our data sets against a series of databases that contain accepted names, synonyms with links to the respective accepted names and suggestions for grouping contentious taxa in larger complexes. If a taxon name was found in one of the databases the name was accepted, changed into the accepted name in case it was a synonym, or grouped into the respective complex. Once a taxon was found in a database, it would not be included in queries of subsequent databases. However, if the accepted name differed from the original one, the accepted name would be queried through all previous databases again. The results were also controlled visually for consistency. The following databases were used in the same order:

1. Table S2 from (Kahlert *et al.* 2020)
2. The taxon list associated with the OMNIDA software (Lecointe *et al.* 1993)
3. The German list of freshwater organisms (Mauch *et al.* 2017)
4. The diat.barcode database (Rimet *et al.* 2019)
5. The website algaebase.org (Guiry 2020)
6. The global biodiversity information platform (gbif) (GBIF.org 2020)

The harmonization of macro-invertebrate data required less effort, and was achieved with gbif (GBIF.org 2020) through the taxize R package (Scott Chamberlain & Eduard Szocs 2013).

3 What is the optimal taxonmic level?

One result of the last progress review for Get Real (held on the 29.04.2020) was that taxa in the TA should be included on their respective optimal taxonomic level instead of using one level (e.g. Genus) for all. *Oligochaetes*, for example, are usually only determined to subclass level, which should not prevent them to be part of a TA in the case that they are common in a given river type. Thus, the question arises: given a data set, what is the optimal taxonomic level to represent a specific taxon?
To establish the optimal level, we used a hierarchical approach. First, we removed all

observations from Phyla and Classes that were not present in all data sets. We assumed that these represented differences in sampling rather than in communities. That left us with the classes Clitellata (Annelida), Insecta, Malacostraca (Arthropoda), Bivalvia and Gastropoda (Mollusca).

In the following, a higher taxonomic level refers to levels with higher resolution, i.e. species is the highest taxonomic level and kingdom the lowest. For each taxon, we calculated the percentage of observations that are represented at each higher level. For example, 4.12% of observations from the order *Lepidoptera* are at the species level, 74.77% at the genus level, 7.75% at the family level, and 13,35% at the order level. Now given a threshold X, which is to be determined, we would call a taxon optimally represented at a certain taxonomic level if less than X% are represented by higher levels. For example, *Lepidoptera* would be represented on order level if $X > 4{,}12 + 74{,}77 + 7{,}75 = 86{,}64\%$. As there are no theoretical grounds on which to base such a threshold value we searched for noticeable patterns in the data (Figure 1). The most noticeable jump occurs between 85 and 86%. It occurs because for $X > 86$ *Chironomidae* are represented at the family level. Hence, we used 85% as threshold. Observations that were missed by this procedure, e.g. observations of *Chironomidae* at the family level, were included at their respective level.

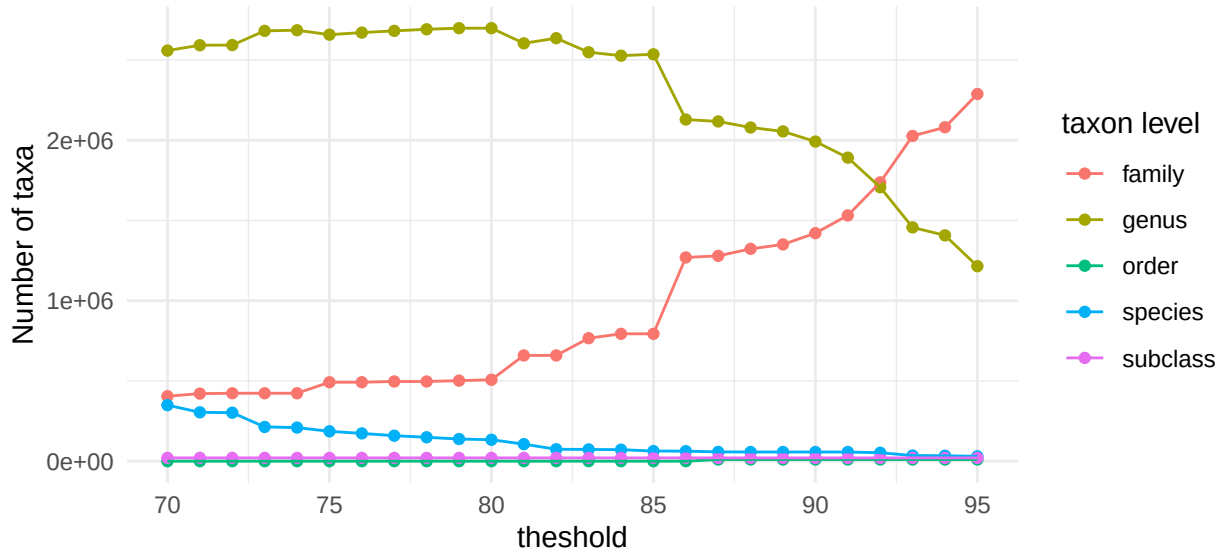

Figure 1: Number of taxa represented at a given taxonomic level as function of the threshold value

For the diatoms, we employed 75% as threshold, because Gomphonema, which is the fourth most common genus in our data set, had 81.43% observations at the species level. The taxonomic resolution was higher than in the macroinvertebrate data and the lowest resolution is the genus level.

Table 1: The ratings for all river types for macro-invertebrates and diatoms

Rating	Taxon	River.Types
high	macro-invertebrates	4, 5, 9, 10, 11, 12, 13, 16
high	diatoms	
medium	macro-invertebrates	1, 2, 3, 8, 14, 15, 18
medium	diatoms	1, 2, 3, 4, 5, 6, 8, 9, 12, 14, 16, 17, 18, 19
low	macro-invertebrates	6, 7, 17, 19, 20
low	diatoms	7, 10, 11, 13, 15, 20

4 Can we represent the stream types with our data?

We determined visually whether a our data set contained enough sampling sites in a given river type to derive meaningful TAs. The degree of representation for river type was graded in a three-tier system: High, medium, and low. A high degree of representation indicates, that we have many sampling locations, which are distributed across the instances of a river type which fall within the countries considered in GetReal. A low degree indicates the opposite, i.e. few and spatially clustered sites. A medium rating implies that we either have many sampling sites, but these only extend over parts of the countries or few sites that extend over most of the countries. The ratings for all river types for both macro-invertebrates and diatoms are shown in table 1.

For each river type x taxon combination, you can find maps with the associated sampling sites for macro-invertebrates and for diatoms.
Further analyses were conducted for all stream types with a high or medium degree of representation. More information on the river types is available in Lyche Solheim *et al.* (2019). In general, we have fewer sampling sites for diatoms than for macro-invertebrates which entails that the representation of stream types is generally lower.

5 What is a typical Assemblage?

We derived TAs based on a rule that considered:

1. The probability of site x belonging to stream type z given that species y is present (a measure of specificity, henceforth **A**)

2. The probability of species y being present given that site x belongs to stream type z (a measure of commonness, henceforth **B**)

3. The Species Indicator Value

The Species Indicator Value (Dufrêne & Legendre 1997; Cáceres & Legendre 2009) is the

weighted product of **A** and **B** (see Equation (1))

$$\sqrt{Indval} = \sqrt{A_g \times B} = \sqrt{\frac{\frac{n_p}{N_p}}{\Sigma_{k=1}^{K} \frac{n_k}{N_k}} \times \frac{n_p}{N_p}} \tag{1}$$

where N_p is the number of sites that belong to river type p and n_p the number of occurrences of the focal species in sites of type p. K is the number of river types. **A** is weighted by the total number of occurrences to account for unequal sample sizes. The statistical significance of the Indicator Value can be assessed with permutation-based pseudo-p-values, which we did with 999 permutations. Here, we are not interested in indicator species for each community, but TAs. Hence, simply continuing with those species that have a pseudo-p-value below some significance level would not serve our purpose. A species that occurs at each site, across all stream types, highlights the difference: while it would not be indicative of any stream type (low specificity) it should be part of each TA. Hence, we need additional criteria to derive the TAs which can be based on **A**, **B**, and the pseudo-p-value of the indicator value. We used the following rules:
For macro-invertebrates:
Species were considered typical if $\mathbf{B} > 0.25$ or $\mathbf{B} > 0.20$ and $p < 0.05$ or $\mathbf{A} > 0.80$
Genera were considered typical if $\mathbf{B} > 0.50$ or $\mathbf{B} > 0.33$ and $p < 0.05$ or $\mathbf{A} > 0.95$
Families were considered typical if $\mathbf{B} > 0.95$ or $\mathbf{B} > 0.80$ and $p < 0.01$ or $\mathbf{A} > 0.99$

For diatoms:
Species where considered typical if $\mathbf{B} > 0.4$ or $\mathbf{B} > 0.3$ and $p < 0.05$ or $\mathbf{A} > 0.7$.
Genera where considered typical if $\mathbf{B} > 0.8$ or $\mathbf{B} > 0.6$ and $p < 0.05$ or $\mathbf{A} > 0.95$.

Note that there was no we did not systematically optimize these thresholds. Such procedures would require optimization criteria, but we are not aware of a criterion that would work in this context. We acknowledge that TA could be (i) very similar in composition or (ii) harbor strongly differing numbers of taxa. Thus, parametrizing the rules in a way that would (i) maximize dissimilarity between assemblages or (ii) maximize the mean assemblage richness would not lead to what we consider a typical assemblage. It would be possible to try a cross-validation-type approach where each taxon is scored based on the number of random-site-subsets it is included in, but such an approach would also entail making essentially arbitrary numerical assumptions. We think the use of subjectively defined thresholds is justified, as long as they are clearly and openly communicated, to be what we define as "typical assemblages".

However, we conducted a sensitivity analysis to see how much varying the parameters of the rules would alter the results. We altered the threshold values of **A** and **B**. The rules above contain two distinct **B** Threshold: B_1 which does not consider the pseudo-p-value ($\mathbf{B} > 0.25$ for macro-invertebrate species and $\mathbf{B} > 0.40$ for diatom species) and B_2 which does take the pseudo-p-value into account ($p < 0.05$ and $\mathbf{B} > 0.2$ macro-invertebrate species and $\mathbf{B} > 0.30$

for diatom species). In the following simulations, the B_2 was always taken to be 25% below B_1. Henceforth, when referring to the threshold for B, we refer to B_1. For species, we varied the threshold for B in ten steps between 0.10 and 0.75 and that for **A** in ten steps between 0.5 and 1.0. For lower taxonomic levels these thresholds were raised. For genera, the threshold values of **A** and **B** were raised by a factor of 1.25 and 2 respectively. All levels family and and lower taxonomic levels were grouped in "families or lower" (fol). For fol, the thresholds were raised by factors of 1.5 and 3 respectively. The taxon richness and uniqueness scores of each TA were computed for all 100 combinations of these parameters and each taxonomic level. Please note that results are only shown and discussed for the non-redundant TAs (see section 6). Taxa richness decreased with increasing **A** and **B** threshold in macro-invertebrates and diatoms (Figure 2A and Figure 3A), while the uniqueness scores increased with **B** thresholds but decreased with **A** thresholds ((Figure 2B and Figure 3B)). Uniqueness scores decreased noticeably with very high **A** thresholds (> 0.9), indicating that taxa that are specific to certain river types are an important driver of TA differentiation. Note that graphs are only shown for all taxa levels combined. Plots for each taxon level separately are available for macro-invertebrates. However, the general patterns visible in Figure 2 and Figure 3, hold for them as well.

Species richness decreased with increasing A and B threshold (Figure 1 and Figure 2), while the uniqueness scores increased with B thresholds but decreased with A thresholds (Figure 3 and Figure 4). The rate of change in species richness along gradients in A and B threshold differed markedly between TAs but seemed to be correlated with overall species richness, i.e. more species-rich TA lost species more quickly than less species-rich ones. The TAs of RT03 and RT16 serve as examples at both extremes of our data set. Uniqueness scores decreased noticeably with very high A thresholds (> 0.9), indicating that taxa that are specific to certain river types are an important driver of TA differentiation.

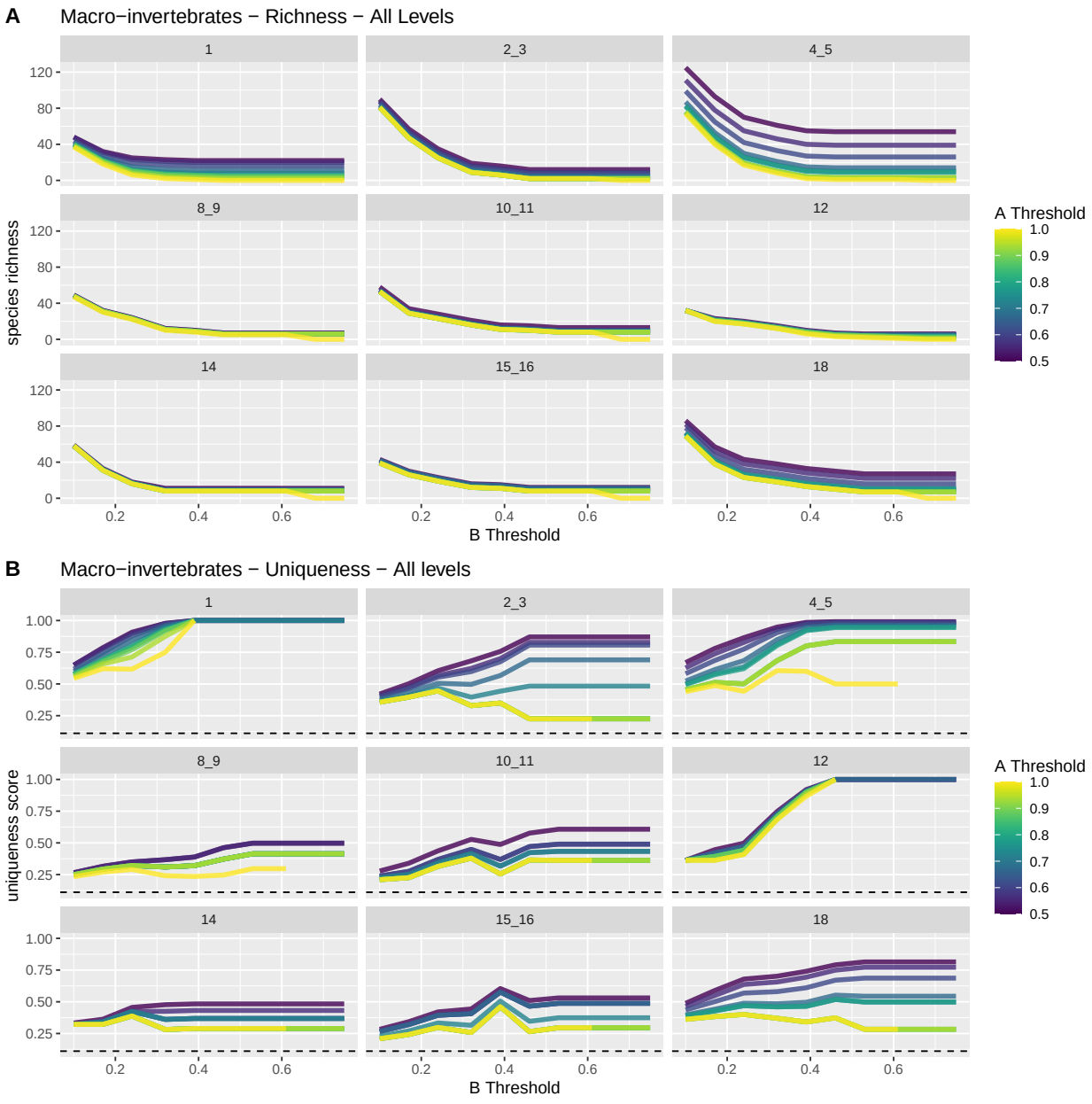

Figure 2: Changes in taxon richness along a changing B threshold. Line color indicates the A threshold

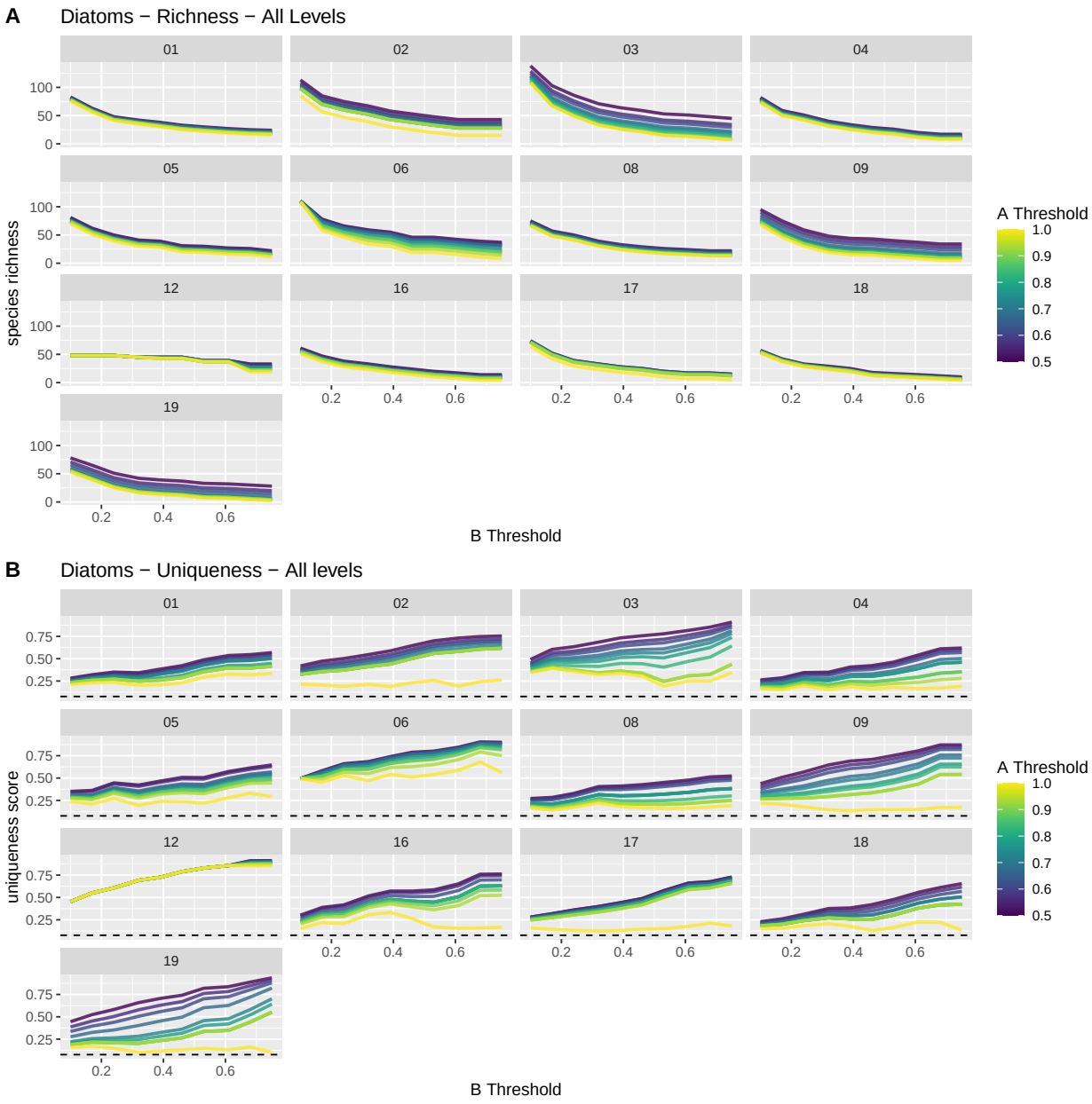

Figure 3: Changes in uniqueness score along a changing B threshold. Line color indicates the A threshold

Table 2: Overlap between different typical assemblages.

River.Type	Macroinvertebrate	Diatoms
RT1	RT2+4 (25%)	RT2 (62.9%)
RT2	RT3 (88.2%)	RT1 (46.8%)
RT3	RT2 (68.2%)	RT2 (48.7%)
RT4	RT3 (45.8%)	RT2 (74.1%)
RT5	RT4 (83.3%)	RT2 (62.5%)
RT6		RT12 (38.3%)
RT8	RT10 (77.8%)	RT2 (71.4%)
RT9	RT8 (76.5%)	RT5+8 (56%)
RT10	RT11+18 (65.2%)	
RT11	RT10 (88.2%)	
RT12	RT9 (50.0%)	RT6 (40.9%)
RT13	RT2 (87.5%)	
RT14	RT16+18 (69.2%)	
RT15	RT16 (85.7%)	
RT16	RT9+10+11+15 (57.1%)	RT18 (56.5%)
RT17		RT1 (63%)
RT18	RT10 (55.6%)	RT1 (71.4%)
RT19		RT17 (68.4%)

6 Redundancy between typical assemblages

We assessed to which degree the different TAs overlap (Table 2). The degree of overlap is the percentage of taxa in a TA that is also present in the most similar (largest overlap) TA. Again, choosing a threshold above which we consider two assemblages to be redundant is somewhat arbitrary. We proceeded with 75% but are open to other suggestions. This threshold leads to five redundant assemblages in macroinvertebrates and none in diatoms. Of the redundant TAs most belong to two river types that only differ in river size: RT02 and 03, 04 and 05, 08 and 09, as well as 10 and 11. The only exception is the combination of RT15 and 16. Both are high altitude river types that occur mainly in southern Europe, which differentiates them from the northern high altitude rivers in RT14. RT13 is also redundant with RT02 and 03 however joining it with these two river types led to a drastically reduced number of taxa in the TA, when compared to that of the combined river type RT02_03. Since RT13 represents an exceedingly rare river type we decided to omit it from the analysis and proceed with RT02_03 instead of RT02_03_13. The new TAs resulted in overall lower degrees of overlap, none of which exceeds the 75% threshold. The largest overlaps were between RT8_9 and RT10_11, with 70%.

7 Notes for Traits

CWM with B value then RR-VGLM (glm RDA)

References

Cáceres, M.D. & Legendre, P. (2009). Associations between species and groups of sites: inindices and statistical inference. *Ecology*, 90, 3566–3574.

Dufrêne, M. & Legendre, P. (1997). Species Assemblages and Indicator Species: The need for a flexible asymmetrical Approach. *Ecological Monographs*, 67, 345–366.

GBIF.org. (2020). *GBIF home page*.

Guiry, G.M., M. D. & Guiry. (2020). *AlgaeBase. World-wide electronic publication, national university of ireland, galway.*

Kahlert, M., Rühland, K.M., Lavoie, I., Keck, F., Saulnier-Talbot, E. & Bogan, D. *et al.* (2020). Biodiversity patterns of Arctic diatom assemblages in lakes and streams: Current reference conditions and historical context for biomonitoring. *Freshwater Biology*, 1–25.

Lecointe, C., Coste, M. & Prygiel, J. (1993). "Omnidia": Software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, 269, 509–513.

Lee, S.S., Bishop, I.W., Spaulding, S.A., Mitchell, R.M. & Yuan, L.L. (2019). Taxonomic harmonization may reveal a stronger association between diatom assemblages and total phosphorus in large datasets. *Ecological Indicators*, 102, 166–174.

Lyche Solheim, A., Austnes, K., Globevnik, L., Kristensen, P., Moe, J. & Persson, J. *et al.* (2019). A new broad typology for rivers and lakes in Europe: Development and application for large-scale environmental assessments. *Science of the Total Environment*, 697, 134043.

Mauch, E., Schmedtje, U., Maetze, A. & Fischer, F. (2017). Taxaliste der Gewässerorganismen Deutschlands. *Informationsberichte des Bayerischen Landesamtes für Wasserwirtschaft*, 1.

Rimet, F., Gusev, E., Kahlert, M., Kelly, M.G., Kulikovskiy, M. & Maltsev, Y. *et al.* (2019). Diat.barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, 9, 1–12.

Scott Chamberlain & Eduard Szocs. (2013). Taxize - taxonomic search and retrieval in r. *F1000Research*.