# OUTLINE: Harmonized macroinvertebrate trait database, Aggregation of traits, Trait definitions, Sources

## 1 Introduction

In the age of the Anthropocene invertebrates are exposed to multiple stressors such as chemical pollution, hydro-morphological modifications, invasive species and climate change. Freshwater invertebrates play a crucial role for many ecosystem processes such as carbon cycling and water purification. In addition to their functional role, 10 % of all described species live in freshwaters, most of which being invertebrates, and thus represent a large part of animal biodiversity. Understanding distributions of invertebrate communities and predicting effects of potential stressors is a long time goal of freshwater research. Since the 1970s ecologists have started using organismal traits to group aquatic insect species

Traits are biological characteristics or ecological preferences measurable at the individual organism. They reflect an organisms adaptation to its habitat. Trait based approaches in ecology are based on the habitat template theory, that predicts where environmental conditions are similar trait composition should also converge, even across biogeographic boundaries. Furthermore, several studies suggested that trait variability is lesser on larger geographical scales than taxonomic variability. Hence, trait based approaches can be a suitable tool for comparing the effect of various environmental stressors to invertebrate communities on large scales or across regions. Consequently, studies have been carried out that e.g. examine the relationship between climate and freshwater community assemblages using species traits. Many of such studies combine information from several trait databases. For example, Brown et al. used in their study traits from invertebrate trait databases from Europe, North America and New Zealand to investigate the effect of decreasing glacier cover to river ecosystems.

In fact, invertebrate trait data are increasingly available. Over the last decades aquatic ecologists have compiled comprehensive databases on invertebrate traits for different regions (see table 1 for examples). However, researchers face various problems when they need to synthesize information from multiple invertebrate trait databases. i) The use of inconsistent terminology across studies (see Schmera et al. 2015 for a comprehensive discussion). For example, some studies used the term "trait" to describe a general organismal property like "generations per year" (Statzner et al. 1997, Usseglio-Polatera et al. 2000), while in other studies this term was related to categories like "bi/multivoltine" (Haybach et al. 2004, Vieira et al. 2006). Here, we follow the proposal of Schmera et al. (2015) and use the term *trait* for a morphological, physiological, or phenomenological feature measurable at the individual organism (e.g. tegument, gills, etc.) and the term *grouping feature* to describe a general property of related organismal traits (e.g. respiration). ii) Invertebrate trait databases from different regions are not standardized. Often, the same grouping features are categorized using different traits. This is complicated by the use of different measurement scales. For example, trait databases from North America use traditionally a binary coding (i.e. trait is expressed or not), whereas most other trait databases (e.g. Tachet, freshwaterecology, and New Zealand) use fuzzy coding (i.e. trait is expressed to a certain extent by the organism). Hence, transformation of nominal scale to ratio scale or vice versa is required. Furthermore, definitions for some traits differ as well across databases. iii) Taxonomical resolution between databases differs, further complicating the synthesis of trait data from different regions. Some trait databases have recorded information on mixed levels of taxonomical resolution like the North American trait database or the Australian trait database. In contrast, trait information in the freshwaterecology database is entirely recorded on species-level. Using trait information on varying taxonomical levels is only possible when traits are aggregated to the lowest taxonomical level that is shared by all used databases. However, iv) so far studies comparing different ways of aggregating traits are lacking.

Given the problems mentioned above extensive data processing is required before researchers can use multiple invertebrate trait databases for their work. In this paper we examine difficulties that ecologists face when synthesizing trait information from different invertebrate trait databases. We explore the effect of different decisions researches have to make when working with invertebrate trait data from several sources, involving trait harmonization, handling different codings, normalization, and aggregation of traits. Therefore, we harmonized six grouping features of different trait databases from four regions and aggregated the trait information to family-level. We discuss the harmonization and show the effect of different ways of aggregating traits (inter alia Problem of different coding styles (fuzzy vs binary)). We also present an overview of differences in trait definitions among databases. Finally, our paper compares the references for the trait information that were specified in the trait databases we used.

# 2 Methods

## 2.1 Description of harmonized trait databases

The harmonized databases are using the available information on aquatic invertebrate traits for the regions Europe, North America, Australia, and New Zealand. Due to the different number and identity of grouping features in each database, the following six grouping features were chosen for this study: locomotion, feeding mode, respiration, voltinism, size, and body form. Trait information for Australia and New Zealand were retrieved from a single database, respectively. For Europe we gathered trait information from the freshwaterecology trait database (`https://www.freshwaterecology.info/`) and complemented where possible missing information with the Tachet trait database (Usseglio-Polatera et al. 2000). North American invertebrate traits were retrieved from Laura Twardochleb and complemented where possible by trait information from Vieira et al (Vieira et al. 2006). From now on, if we use the term European or North American trait database we refer to the combined databases unless explicitly stated otherwise. We used all available information on invertebrates in the databases but restricted our analysis to those taxa that have complete trait profiles for the grouping features mentioned earlier. We are aware that imputation methods exist, which infer missing information for traits by interpolating from related traits. However, by using only complete data we were able to evaluate the taxonomical coverage per order within the databases. We consider this a helpful information for researchers who strive to fill data gaps. Table 1 gives an overview of the used databases. In the following paragraphs, we describe the data processing steps required to establish a harmonized invertebrate trait database.

Table 1: Overview of trait databases.

| Region | Coding of trait states | Reference |
|---|---|---|
| Europe | Largely fuzzy | Schmidt-Kloiber and Hering 2015 |
| Central Europe | Fuzzy coded | Usseglio-Polatera et al. 2000 |
| North America | Largely binary | Vieira et al. 2006 |
| North America | Largely binary | cite Laura Twardochleb |
| Australia | Binary & fuzzy coding | Kefford et al. 2019 |
| New Zealand | Fuzzy coded | |

## 2.2 Normalization & data conversion

Establishing a harmonized database required traits on the same measurement scale to enable comparability. However, in the used databases traits varied within and among databases in terms of measurement scale. Traits were either fuzzy coded, binary coded or to a small extent coded as continuous variable (e.g. size). Fuzzy codes capture the variation (i.e. temporal or spatial) in a trait and are represented by affinity scores (sometimes termed membership state) which express a preference of a taxon to a certain trait. Usually, these scores range from 0 representing no affinity to an arbitrary maximum value which represents high affinity. This type of coding uses the ratio scale and not the ordinal scale as one might intuitively assume. In fact, the intention of the developers of this coding system was to convert affinity scores to percentages per trait, e.g. for a grouping feature with four traits affinity scores of *5, 6, 2, 2* are equal to *1/3, 2/5, 2/15, 2/15* or *33.3 %, 40 %, 13.3 %, 13.3 %* (Chevenet, Dolédec, and Chessel 1994). We used fuzzy coded traits for establishing our harmonized databases where possible, otherwise binary traits. Categorical and continuous traits across all used databases were converted into binary traits. Implicitly, we assumed for binary variables that a value of 1 for a particular trait corresponded to the maximum affinity of a taxon for that particular trait.

Two databases needed further data processing. Firstly, the Australian database which is a collection of seven trait databases. Thus, several grouping features occur multiple times but with traits that have different types of codings and ranges. For example, body size occurs as continuous variable, with traits that are fuzzy coded, and with traits that are binary coded. The same traits originating from different sub-databases in the Australian trait database were allocated. To enable allocation of fuzzy coded and binary traits we applied a range normalization by dividing each fuzzy coded trait by its potential maximum value. Consequently, the values of each trait were converted to a range from 0 to 1.

Secondly, the North American trait database which only contained traits on the nominal scale. We first converted nominal traits into binary traits. As a second approach we converted nominal traits for taxa on species and genus-level into "pseudo" fuzzy codes by calculating the percentage of occurrence for each trait per genera and grouping feature. Both approaches were compared regarding the trait values they yielded. We refer to the version with binary traits as *NOA_bin* and to the version with pseudo fuzzy codes as *NOA_fuzzy*.

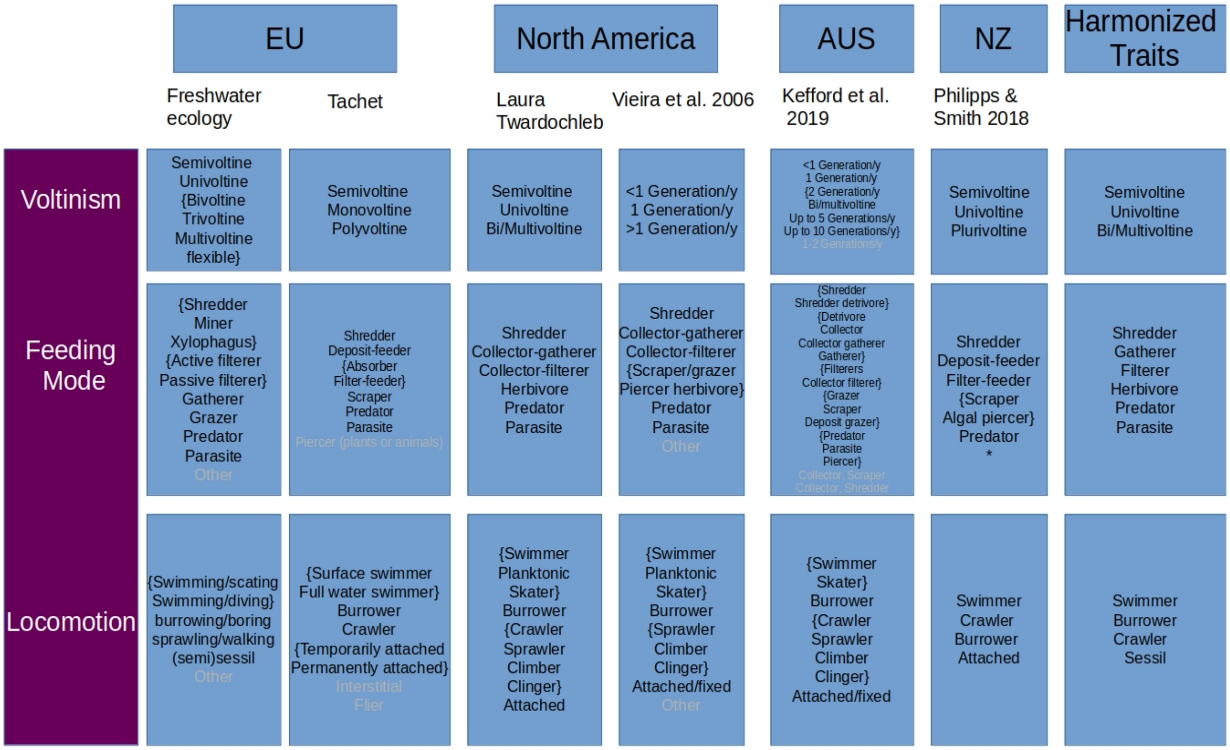
The Australian, Tachet, and the North American trait databases contained duplicate taxa entries - either on species, genus or family-level - which were amalgamated. The duplicate entries in the Australian trait database originated from the different sub-databases. As a result, many of the duplicate entries complemented each other (e.g. for a given trait one duplicate contained a value, the other not). In those cases where several different values for a trait occurred, the average trait value was calculated. The NOA_bin database only contained 0 and 1 as entries. Hence, for duplicates the maximum value was taken. For the tachet trait database duplicate entries were allocated using the mean if they contained contrasting information on the same trait.

Finally, every trait within each grouping feature was standardized by the sum of all traits within this particular grouping feature per taxon to convert traits to percentages per trait.

## 2.3 Harmonization of traits and taxa

Harmonization of traits means amalgamating several similar traits into a single trait. It has to be undertaken when grouping features from different sources (e.g. different regions) are used and the grouping features do not contain the same traits. For example, in our study for the grouping feature locomotion the lowest number of traits that occur across all databases was in the New Zealand trait database ("Swimmer", "Burrower", "Crawler", "Sessil"). Hence, in all other databases locomotion traits have been allocated into these four traits (see 1 and 2).

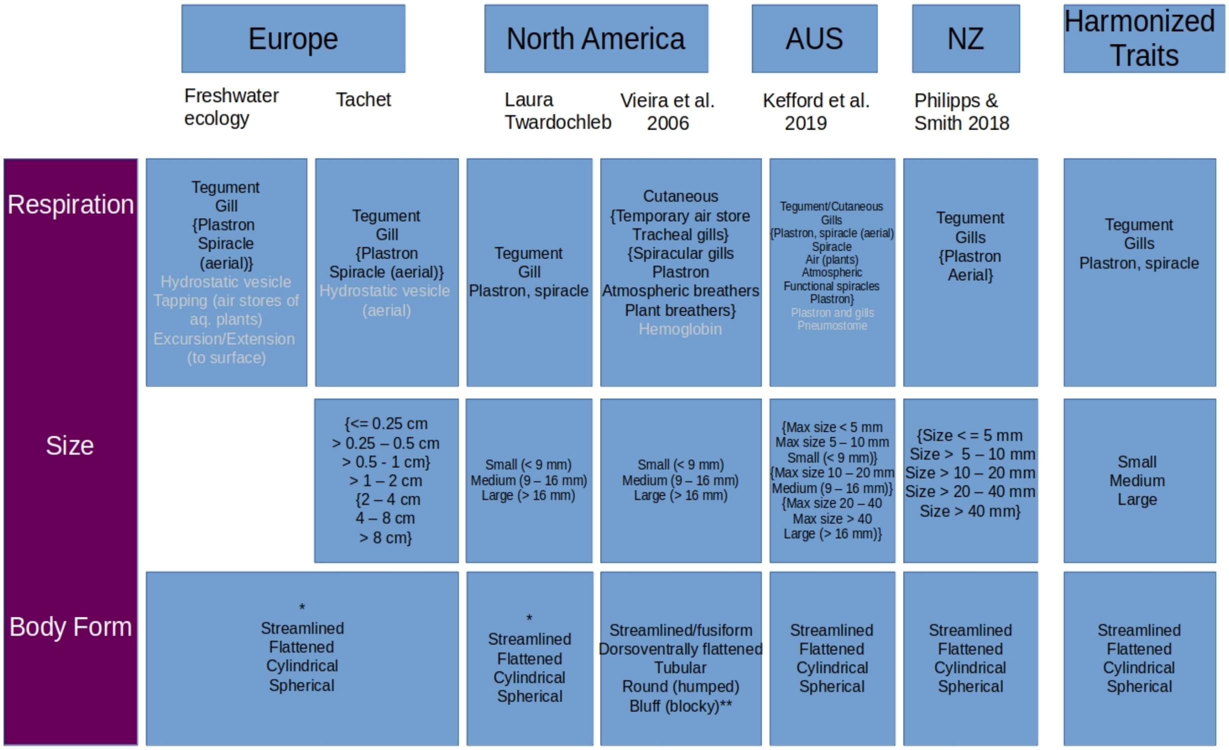
Grouping features that differed in their traits among databases have been harmonised by condensing the traits in such a way that in the end the same grouping features in all databases consisted of the same traits. Thereby, traits were amalgamated based on ecological knowledge or expert judgment. Our approach of harmonizing the traits in the used databases is outlined in figures 1 and 2.



| | EU | | North America | | AUS | NZ | Harmonized Traits |
|---|---|---|---|---|---|---|---|
| | Freshwater ecology | Tachet | Laura Twardochleb | Vieira et al. 2006 | Kefford et al. 2019 | Philipps & Smith 2018 | |
| Voltinism | Semivoltine<br>Univoltine<br>{Bivoltine<br>Trivoltine<br>Multivoltine<br>flexible} | Semivoltine<br>Monovoltine<br>Polyvoltine | Semivoltine<br>Univoltine<br>Bi/Multivoltine | <1 Generation/y<br>1 Generation/y<br>>1 Generation/y | <1 Generation/y<br>1 Generation/y<br>{2 Generation/y<br>Bi/multivoltine<br>Up to 5 Generations/y<br>Up to 10 Generations/y}<br>1-2 Genrations/y | Semivoltine<br>Univoltine<br>Plurivoltine | Semivoltine<br>Univoltine<br>Bi/Multivoltine |
| Feeding Mode | {Shredder<br>Miner<br>Xylophagus}<br>{Active filterer<br>Passive filterer}<br>Gatherer<br>Grazer<br>Predator<br>Parasite<br>Other | Shredder<br>Deposit-feeder<br>{Absorber<br>Filter-feeder}<br>Scraper<br>Predator<br>Parasite<br>Piercer (plants or animals) | Shredder<br>Collector-gatherer<br>Collector-filterer<br>Herbivore<br>Predator<br>Parasite | Shredder<br>Collector-gatherer<br>Collector-filterer<br>{Scraper/grazer<br>Piercer herbivore}<br>Predator<br>Parasite<br>Other | {Shredder<br>Shredder detrivore}<br>{Detrivore<br>Collector<br>Collector gatherer<br>Gatherer}<br>{Filterers<br>Collector filterer}<br>{Grazer<br>Scraper<br>Deposit grazer}<br>{Predator<br>Parasite<br>Piercer}<br>Collector, Scraper<br>Collector, Shredder | Shredder<br>Deposit-feeder<br>Filter-feeder<br>{Scraper<br>Algal piercer}<br>Predator<br>* | Shredder<br>Gatherer<br>Filterer<br>Herbivore<br>Predator<br>Parasite |
| Locomotion | {Swimming/scating<br>Swimming/diving}<br>burrowing/boring<br>sprawling/walking<br>(semi)sessil<br>Other | {Surface swimmer<br>Full water swimmer}<br>Burrower<br>Crawler<br>{Temporarily attached<br>Permanently attached}<br>Interstitial<br>Flier | {Swimmer<br>Planktonic<br>Skater}<br>Burrower<br>{Crawler<br>Sprawler<br>Climber<br>Clinger}<br>Attached | {Swimmer<br>Planktonic<br>Skater}<br>Burrower<br>{Sprawler<br>Climber<br>Clinger}<br>Attached/fixed<br>Other | {Swimmer<br>Skater}<br>Burrower<br>{Crawler<br>Sprawler<br>Climber<br>Clinger}<br>Attached/fixed | Swimmer<br>Crawler<br>Burrower<br>Attached | Swimmer<br>Burrower<br>Crawler<br>Sessil |

Figure 1: Proposed harmonization scheme for the grouping features voltinism, feeding mode and locomotion. Shown are all traits for the used grouping features in the investigated trait databases and the harmonized traits in the end. Traits in curly brackets were harmonized to one trait. Traits highlighted in Grey were omitted.

** Trait parasite was not available in New Zealand trait database.*



| | Europe | | North America | | AUS | NZ | Harmonized Traits |
|---|---|---|---|---|---|---|---|
| | Freshwater ecology | Tachet | Laura Twardochleb | Vieira et al. 2006 | Kefford et al. 2019 | Philipps & Smith 2018 | |
| Respiration | Tegument<br>Gill<br>{Plastron<br>Spiracle<br>(aerial)}<br>Hydrostatic vesicle<br>Tapping (air stores of aq. plants)<br>Excursion/Extension (to surface) | Tegument<br>Gill<br>{Plastron<br>Spiracle (aerial)}<br>Hydrostatic vesicle (aerial) | Tegument<br>Gill<br>Plastron, spiracle | Cutaneous<br>{Temporary air store<br>Tracheal gills}<br>{Spiracular gills<br>Plastron<br>Atmospheric breathers<br>Plant breathers}<br>Hemoglobin | Tegument/Cutaneous<br>Gills<br>{Plastron, spiracle (aerial)<br>Spiracle<br>Air (plants)<br>Atmospheric<br>Functional spiracles<br>Plastron}<br>Plastron and gills<br>Pneumostome | Tegument<br>Gills<br>{Plastron<br>Aerial} | Tegument<br>Gills<br>Plastron, spiracle |
| Size | | {<= 0.25 cm<br>> 0.25 – 0.5 cm<br>> 0.5 - 1 cm}<br>> 1 – 2 cm<br>{2 – 4 cm<br>4 – 8 cm<br>> 8 cm} | Small (< 9 mm)<br>Medium (9 – 16 mm)<br>Large (> 16 mm) | Small (< 9 mm)<br>Medium (9 – 16 mm)<br>Large (> 16 mm) | {Max size < 5 mm<br>Max size 5 – 10 mm<br>Small (< 9 mm)}<br>{Max size 10 – 20 mm<br>Medium (9 – 16 mm)}<br>{Max size 20 – 40<br>Max size > 40<br>Large (> 16 mm)} | {Size < = 5 mm<br>Size > 5 – 10 mm<br>Size > 10 – 20 mm<br>Size > 20 – 40 mm<br>Size > 40 mm} | Small<br>Medium<br>Large |
| Body Form | *<br>Streamlined<br>Flattened<br>Cylindrical<br>Spherical | | *<br>Streamlined<br>Flattened<br>Cylindrical<br>Spherical | Streamlined/fusiform<br>Dorsoventrally flattened<br>Tubular<br>Round (humped)<br>Bluff (blocky)** | Streamlined<br>Flattened<br>Cylindrical<br>Spherical | Streamlined<br>Flattened<br>Cylindrical<br>Spherical | Streamlined<br>Flattened<br>Cylindrical<br>Spherical |

Figure 2: Proposed harmonization scheme for the grouping features respiration, size and body form.
** Body form information provided by Philippe Usseglio-Polatera.*
*** Bluff(blocky) taxa have been reclassified by Philippe Usseglio-Polatera using the traits streamlined, flattened, cylindrical and spherical.*

Not only the categorization of grouping features but also the definitions of individual traits vary between databases, complicating harmonization.

## 2.4 Aggregation of traits

Traits in the processed and harmonized trait databases were aggregated using three approaches. I) taxa on species-level and genus-level were stepwise aggregated to the family-level by initially allocating them at the genus-level using the median. Then all traits were aggregated to family-level by using the mode. In cases where it was not possible to take the mode, e.g. only distinct values, multiple duplicates, or multiple duplicates and distinct values occurred the mean was taken. Hereafter, we abbreviate this aggregation type as *stepwise_agg*. II) we directly aggregated taxa to family level using the median. We denote this aggregation as *direct_agg*. III) taxa were aggregated using a weighted approach, denoted as *weighted_agg*. The weights were determined as the ratio of how many taxa on species or genus-level were initially present per genera compared to how many taxa on species or genus-level were present after selecting only taxa with complete trait profiles. After determining the weights, taxa on species and genus-level were were aggregated to genus-level by multiplying their trait values by their respective weights. Then the weighted trait values were summed up per family for each trait.

The resulting aggregated trait values were compared to trait values assigned at family-level by experts. Trait assignments on family-level existed for the Australian database and the North American database, but only for a limited subset of grouping features and taxa. For the Australian database we could compare aggregated trait values with assigned trait values for the grouping features feeding mode and size. For the North American database we compared aggregated trait values with assigned trait values for the grouping features feeding mode, respiration, size, voltinism and locomotion, albeit only for aquatic insects.

# 3 Results

## 3.1 Taxonomical coverage

Regarding the taxonomical coverage the New Zealand database has, as expected, the smallest taxon pool. In total 492 taxa are covered by this database. Thereof, 404 taxa resolved on species-level, 47 taxa on genus-level and 27 taxa on family-level. The remaining entries are on a lower taxonomical resolution. 73 % of taxa in the New Zealand database belong to the group of aquatic insects. The largest taxon pool is spanned by the European trait database, with 4224 taxa of 76 different orders. 48 % of the taxa in this database belong to the group of aquatic insects. The European database is mostly on the highest taxonomical resolution possible, with 3953 taxa on species-level (approximately 93,6 %). 253 entries are resolved on genus-level and 18 entries on family-level. The Australian database has 1404 taxa of 64 orders. 52 % of the taxa covered are aquatic insects. 564 taxa are resolved on species-level, 578 on genus-level and 260 on family-level. The North American trait database contained trait information for 3542 taxa of 42 different orders, although 63 % of the taxa in the database belong to the aquatic insects. 2142 entries are on species-level, 1074 on genus-level and 50 on family-level.

## 3.2 Completeness of trait information

The percentage of entries with available information for the individual grouping features in the Australian, European and North American trait databases varied between 5 % and 99 %. By contrast, the New Zealand trait database contained complete trait information for 99 % to 100 % of their entries for the individual grouping features (2). The greatest data gap was for the grouping feature body form where information was only present for 7 % of entries in the Australian and European database, and 26 % of entries in the North American database. Selecting only taxa with complete trait profiles for the six grouping features and resolved at least at family-level lead to the omission of many taxa in all databases except for the New Zealand database. Out of the 21 orders that the New Zealand database covers, taxa of 20 orders remained after the selection process. For most orders, all families initially included were also included after selecting taxa with complete trait profiles (table 3). By contrast, in the Australian database, only taxa from five orders (Ephemeroptera, Megaloptera, Odonata, Plecoptera, and Trichoptera) remained. Within these five orders, there was none where all families included in the database contained complete trait profiles. In the North American database, taxa of 18 orders had complete trait profiles, in the European database 9 orders.

Table 2: How many entries have information for the individual grouping features per database.

| Database | Grouping feature | Trait covered [%] |
|---|---|---:|
| Australia | Body form | 5.00 |
| Australia | Feeding mode | 99.00 |
| Australia | Locomotion | 42.00 |
| Australia | Respiration | 70.00 |
| Australia | Size | 78.00 |
| Australia | Voltinism | 49.00 |
| Europe | Body form | 7.00 |
| Europe | Feeding mode | 65.00 |
| Europe | Locomotion | 33.00 |
| Europe | Respiration | 56.00 |
| Europe | Size | 11.00 |
| Europe | Voltinism | 24.00 |
| New Zealand | Body form | 100.00 |
| New Zealand | Feeding mode | 99.00 |
| New Zealand | Locomotion | 99.00 |
| New Zealand | Respiration | 100.00 |
| New Zealand | Size | 100.00 |
| New Zealand | Voltinism | 100.00 |
| North America | Body form | 26.00 |
| North America | Feeding mode | 61.00 |
| North America | Locomotion | 51.00 |
| North America | Respiration | 44.00 |
| North America | Size | 75.00 |
| North America | Voltinism | 47.00 |

Table 3: Proportion of families per order that remain after only taxa with complete trait profiles have been selected from the total number of distinctive families in the databases.

| Order | Australia | New Zealand | Europe | North America |
|---|---|---|---|---|
| Amphipoda | | 100.00 | 37.50 | 10.00 |
| Anthoathecata | | 100.00 | | |
| Architaenioglossa | | | | 100.00 |
| Arhynchobdellida | | | | 100.00 |
| Branchiopoda | | | | 100.00 |
| Coleoptera | | 88.89 | 83.33 | 15.62 |
| Cycloneritida | | | | 100.00 |
| Decapoda | | 100.00 | | |
| Diptera | | 100.00 | 46.15 | 21.62 |
| Ephemeroptera | 50.00 | 100.00 | 66.67 | 73.91 |
| Hemiptera | | 50.00 | 63.64 | 47.06 |
| Hexapoda | | 100.00 | | |
| Lepidoptera | | 100.00 | | |
| Littorinimorpha | | | | 66.67 |
| Mecoptera | | 100.00 | | |
| Megaloptera | 50.00 | 100.00 | | 50.00 |
| Mollusca | | 100.00 | | |
| Mysida | | 100.00 | | |
| Nemertea | | 100.00 | | |
| Neuroptera | | 100.00 | | |
| Odonata | 6.90 | 100.00 | 100.00 | 50.00 |
| Oligochaeta | | 100.00 | | |
| Onychura | | | | 100.00 |
| Plecoptera | 25.00 | 100.00 | 100.00 | 88.89 |
| Rhynchobdellida | | 100.00 | | 50.00 |
| Spinicaudata | | | | 100.00 |
| Tanaidacea | | 100.00 | | |
| Trichoptera | 52.17 | 100.00 | 95.45 | 71.43 |
| Venerida | | | 100.00 | 33.33 |

## 3.3 Deviance in trait values

The *stepwise_agg* and *direct_agg* approaches yielded for the majority of taxa the same trait values (table 4) across all grouping features. The two aggregation methods differed the most when applied to the Australian trait database, with 3.52 %

Table 4: Percentage of cases where the *stepwise_agg* and *direct_agg* methods resulted in different trait values on family-level.

| Database | Deviating cases [%] |
|---|---:|
| Australia | 3.52 |
| Europe | 2.17 |
| New Zealand | 1.03 |
| North America | 2.66 |