A workflow to integrate ecological monitoring data from different sources

Jérémy Wicquart, Mishal Gudka, David Obura, Murray Logan, Francis Staub, David Souter, Serge Planes

A workflow to integrate ecological monitoring data from different sources

Jérémy Wicquart[1,2], Mishal Gudka[3], David Obura[3], Murray Logan[4], Francis Staub[5], David Souter[4], Serge Planes[1,2]

[1] PSL Université Paris: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

[2] Laboratoire d'Excellence "CORAIL"

[3] CORDIO East Africa, Mombasa, Kenya

[4] Australian Institute of Marine Science, PMB No. 3, Townsville MC, Townsville, QLD 4810, Australia

[5] International Coral Reef Initiative

Corresponding author: Jérémy Wicquart ; **Corresponding author email**: jeremywicquart@gmail.com

Abstract

Programs and initiatives aiming to protect biodiversity and ecosystems have increased over the last decades in response to their decline. Most of these are based on monitoring data to quantitatively describe trends in biodiversity and ecosystems. The estimation of such trends, at large scales, requires the integration of numerous data from multiple monitoring sites. However, due to the high heterogeneity of data formats and the resulting lack of interoperability, the data integration remains sparsely used and synthetic analyses are often limited to a restricted part of the data available.

Here we propose a workflow, comprising four main steps, from data gathering to quality control, to better integrate ecological monitoring data and to create a synthetic dataset that will make it possible to analyse larger sets of monitoring data, including unpublished data.

The workflow was designed and applied in the production of the *Status of Coral Reefs of the World: 2020* report, where more than two hundred individual datasets were integrated to assess the status and trends of hard coral cover at the global scale. The workflow was applied to two case studies and associated R codes, based on the experience acquired during the production of this report.

The proposed workflow allows for the integration of datasets with different levels of taxonomic and spatial precision, with a high degree of reproducibility. It provides a conceptual and technical framework for the integration of ecological monitoring data, allowing for the estimation of temporal trends in biodiversity and ecosystems or to test ecological hypotheses at larger scales.

Key-words

Data · Homogenization · Integration · Standardization · Aggregation · Ecological synthesis · GCRMN · Meta-analysis

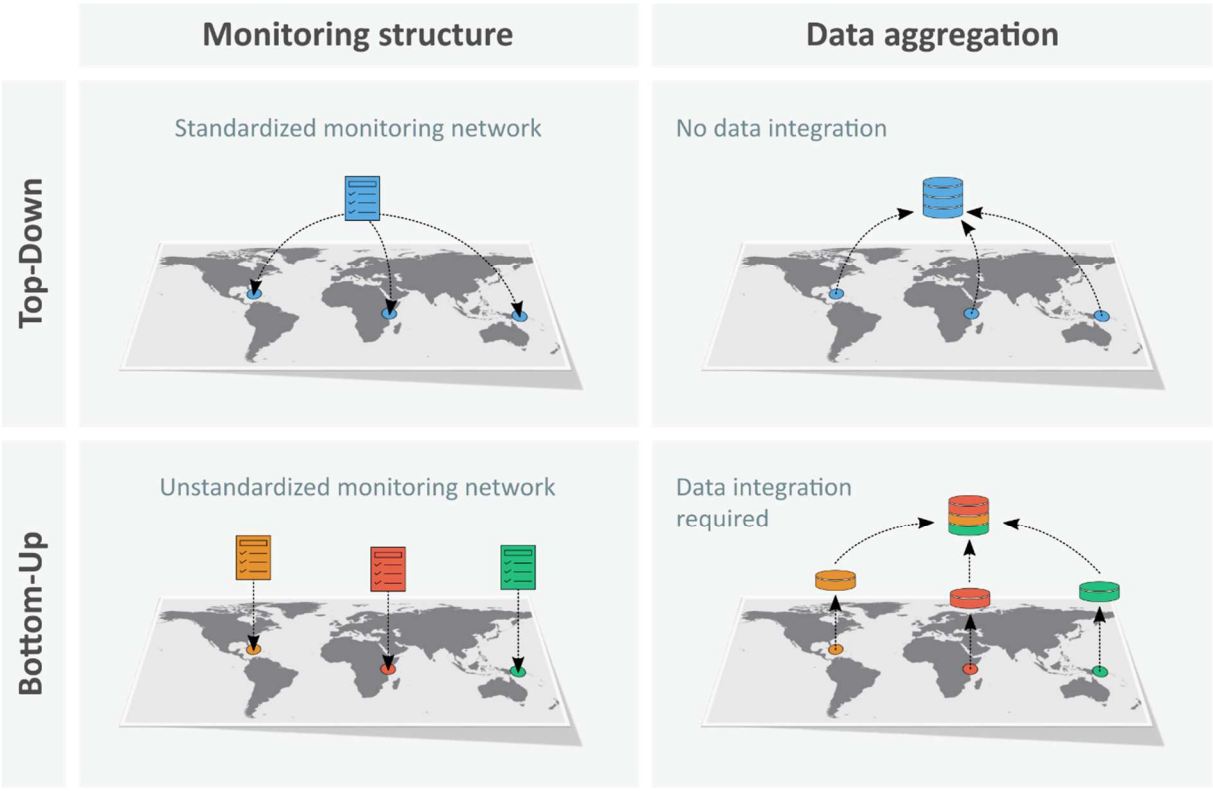
1. Introduction

Global (e.g. Aichi Biodiversity Targets of the Convention for the Biological Diversity - CBD), regional (e.g. Coastal Oceans Research and Development Indian Ocean - CORDIO) and national (e.g. French Coral Reef Initiative) initiatives that seek to protect biodiversity and ecosystems have multiplied over the last two decades. To gauge the success of these different programs in order to inform conservation policies, it is necessary to estimate changes in biodiversity and ecosystems at each of these different scales. Ecological monitoring (see **Table 1** for definitions of mains terms used) is the foundation on which such studies are built, yet, it is typically focused and constrained to local scales. Therefore, in order to assess the broader status and trends of biodiversity and ecosystems, it is necessary to group monitoring sites within a monitoring network (Balmford et al., 2005; Lindenmayer & Likens, 2010; Schmeller et al., 2015; Külh et al., 2020).

Two approaches can be adopted to achieve this objective (Henry et al., 2008). The first is a "top-down" approach and corresponds to a highly standardized monitoring network where all monitoring sites within the network use the same protocol (**Fig. 1**). Monitoring networks that are based on this "top-down" approach are usually found within national frameworks (e.g. US National Coral Reef Monitoring Program - National Oceanic and Atmospheric Administration (NOAA) CoRIS) or within research groups (e.g. Service National Observation (SNO) CORAIL). However, while standardized protocols are used for some monitoring networks (e.g. Hallmann et al., 2017), most monitoring sites are far from being part of a real "top-down" approach due to issues around coordination and standardisation between different research groups who have different objectives, interests, funding streams and capacities. Moreover, the "top-down" approach does not allow for the inclusion of existing monitoring sites that use different protocols. Existing monitoring sites cannot be greatly modified since doing this would lead to a loss of consistency in methodology and hence comparability of data over time, which is one of the main targets of long-term monitoring (Lindenmayer & Likens, 2009). To resolve these issues, a second strategy, the "bottom-up" approach,

69 may be considered (**Fig. 1**). This strategy consists of combining data (hereafter called data
70 integration) acquired from existing monitoring sites that use different methodologies. In contrast to
71 the "top-down" approach, this strategy enables the existing condition of monitoring networks to be
72 considered for data integration, where it would otherwise be impossible.

73



74

75 **Figure 1**. Comparison of monitoring structure and data aggregation between top-down and bottom-
76 up approaches. For the top-down approach, the monitoring network is standardized and is based on
77 a unique protocol and data format. In contrast, for the bottom-up approach, the monitoring network
78 is unstandardized and is based on different protocols and data formats, making data integration
79 necessary for synthetic analyses.

80

81

In contrast to other scientific fields such as physics, oceanography or genetics, ecology is based on a high diversity and heterogeneity of data collection methods and hence data formats (Reichman, Jones, & Schildhauer, 2011; Michener & Jones, 2012; Poisot, Bruneau, Gonzalez, Gravel, & Peres-Neto, 2019), and this, despite the existence of data standards (e.g. DarwinCore (Wieczorek et al., 2012)). This leads to a lack of interoperability between datasets, which represents a major challenge for wider data integration. This is likely one of the main reasons why data integration, which is needed for the "bottom-up" approach, remains poorly developed (Henry et al. (2008) but see Miller, Pacifici, Sanderlin, & Reich, 2019; O'Donnell et al., 2021) outside of large databases such as GBIF (GBIF: The Global Biodiversity Information Facility, 2021) or OBIS (Ocean Biodiversity Information System, OBIS (2021)). An increased use of data integration may lead to a deeper understanding of status and trends in biodiversity and ecosystems, without having to acquire new data (Jones, Schildhauer, Reichman, & Bowers, 2006; Carpenter et al., 2009). Moreover, as emphasized by Borregaard and Hart (2016), data preparation, which includes data integration, is barely considered and reported as part of data analysis and is rarely documented in code associated with published articles. This represents a major issue for transparent science and reproducibility as the data preparation step can also contain errors that other researchers must be able to track. Finally, this also limits the ability for other researchers to rely on existing methods to conduct similar studies.

To address these issues, we propose a workflow which integrates ecological monitoring data from different data sources into a synthetic dataset, which can then be used to perform national, regional or global analyses on the status and trends of the considered ecological metric. We illustrate the proposed workflow by providing an R code template for two case studies inspired from the *Status of Coral Reefs of the World: 2020* report, where 248 datasets from contributors across the world were integrated to estimate the status and trends of hard coral cover at the global scale (Souter et al., 2021).

Table 1. Definitions of main terms used in the article.

Term	Definition
Dataset	A collection of related sets of information that is composed of separate elements (data files) but can be manipulated as a unit by a computer.
Data aggregator	Data analyst responsible for the data integration process.
Data integration	Process of combining, merging, or joining data together, in order to make what were distinct, multiple data objects, into a single, unified data object (Schildhauer, 2018).
Data provider	A person or an institution sharing a dataset for which they have been or are involved in the acquisition of the data contained in the dataset.
Monitoring site	Repetitive measurement of a specified set of variables at one location over an extended period of time (Vos, Meelis, & Ter Keurs, 2000).
Synthetic dataset	A dataset resulting from the integration of multiple existing datasets (Poisot et al., 2016).

2. Workflow

We distinguished four main data sources: databases, data papers, research articles with associated data and unpublished data from data providers. Over the last decades, large databases which gathered data from different monitoring sites, have emerged in ecology, such as ILTER (Vanderbilt & Gaiser, 2017), GBIF (GBIF: The Global Biodiversity Information Facility, 2021) or BioTIME (Dornelas et al., 2018). In addition to these databases, an increasing number of data papers are being published (Shin et al., 2020), extending the availability of monitoring data. However, based on our experience with the *Status of Coral Reefs of the World: 2020* report (Souter et al., 2021), the vast majority of monitoring data remains unpublished, or only partially published, and thus, can only be acquired from direct exchanges with data providers. For this reason, we chose to focus the proposed workflow on the acquisition of unpublished data, while also making it possible to incorporate data from databases, research articles and data papers.

We identified three main approaches which have the potential to yield a synthetic dataset: (1) propose a web-based interface for data entry by data providers (e.g. Chaudhary, Walters, Bever, Hoeksema, & Wilson (2010), Robertson et al. (2014)), (2) ask data providers to reformat their data

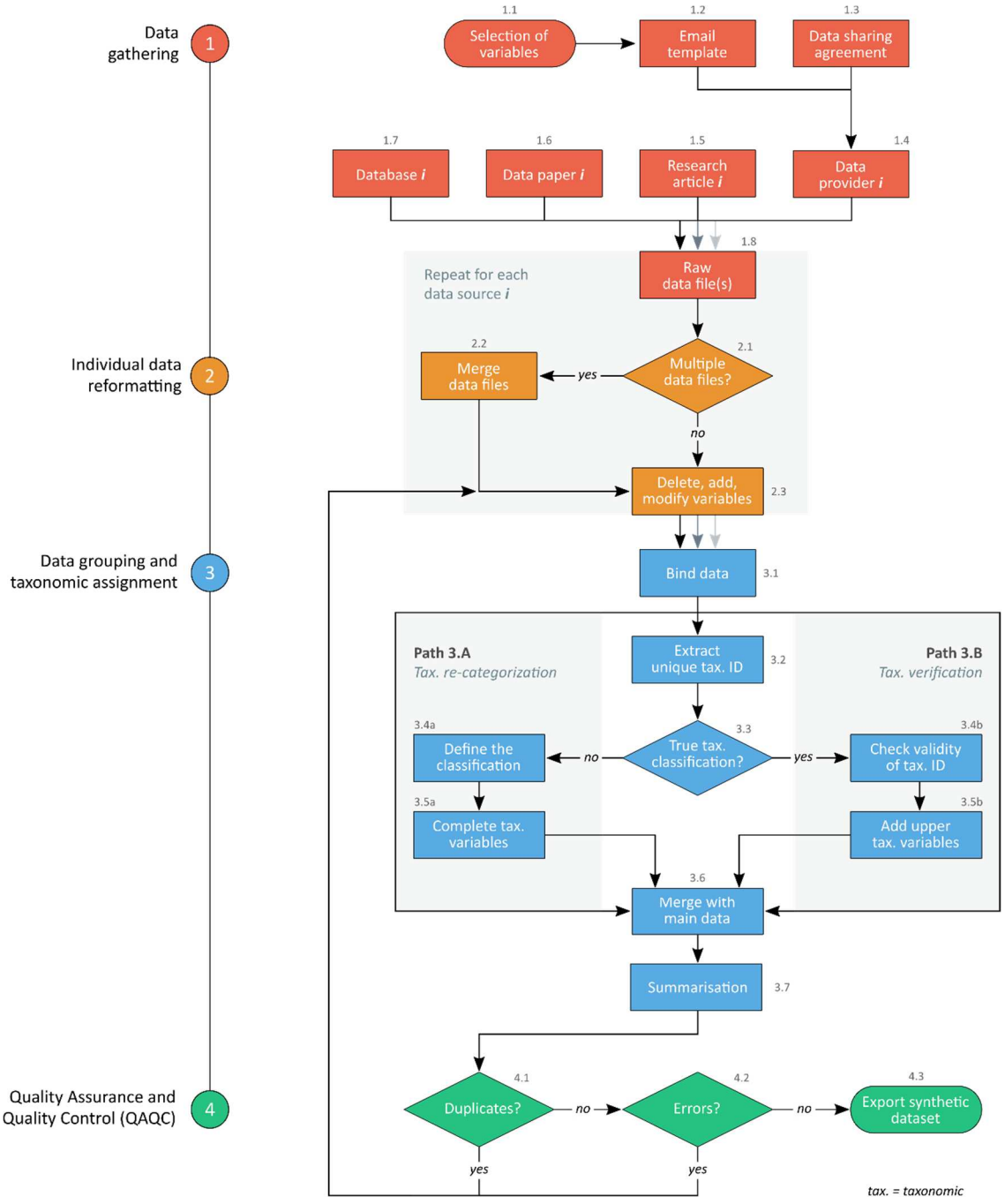
following a given template and (3) collect data from data providers in their original format and centralize the reformatting by a data aggregator. The first approach is particularly adapted for new monitoring networks but not for a “bottom-up” approach, as the entry of historical data can be extremely time consuming. The second approach necessitates data wrangling skills from data providers, as well as time, which can potentially discourage them from contributing. In contrast, centralising the entire data homogenization procedure, allows for greater standardization of homogenization, enables full tracking of changes and biases, and avoids error due to variability in data wrangling skills among data providers. Moreover, the first two approaches are difficult to implement for databases, data papers and research articles, as the associated data are only available in a particular format, which need to be reformatted. For all of these reasons, we chose to build the workflow around the third approach, centralized data reformatting where data is curated by a dedicated data aggregator. This approach could interest both parties involved, as the data provider could benefit from the expertise of the data aggregator on data shared, as well as providing advice on metadata information or potential errors (Costello, Michener, Gahegan, Zhang, & Bourne, 2013).

The workflow (**Fig. 2**) is composed of four main steps: (1) data gathering, (2) individual data reformatting, (3) data grouping and taxonomic assignment and (4) quality assurance and quality control (QAQC). These four different steps are detailed in the following sections and are exemplified by two cases studies and associated R code template.

The workflow was developed with the R software (version 4.1.0, R Core Team (2021)) but it can be transposed into other programming languages (e.g. Python). For R users, we strongly recommend a migration to the “tidyverse” meta-package (Wickham et al., 2019), as it provides a wide range of functions used in data analysis, from importation of the data to their visualisation. In addition to the software, we also recommend using a version control system (e.g. Git) associated to an online collaborative platform (e.g. Github), particularly if the project will be maintained over time and/or if it involves a team of data aggregators. Finally, we highlight that all steps which involve the

150 modification of data, must be done using code and not manually in order to reach the highest
151 possible level of reproducibility and traceability.

152



153

154 **Figure 2.** Workflow from data gathering to exportation of the final synthetic dataset. The diamonds
155 indicate choices, rectangles indicate actions, and ellipses indicate start and end of the workflow.

2.1. Data gathering

The first step in the workflow is the selection of variables that will have to be present in the final synthetic dataset (hereafter "standard variables"; **Fig. 2 - 1.1**). We propose a classification of these variables into six groups: data descriptor (e.g. dataset ID, data source), spatial (e.g. latitude, longitude, depth), temporal (e.g. year, date, hour of sampling), methodological (e.g. length of the transect, number of quadrat), taxonomic (e.g. family, genus, species) and metric (e.g. percentage cover, abundance, size). The metric(s) variable(s) correspond to the response(s) variable(s) while all others correspond to potential explanatory variables or metadata. The variables selected in each of these groups depends on the goal of the project and on the future analyses that will be performed. Among the data descriptor variable group, a variable corresponding to the ID of each dataset must be included in order to ensure the possibility of extracting or to perform sensitivity analyses on the individual datasets. We emphasize that the groups of spatial and taxonomic variables are nested. For example, in the case of spatial variables, we can have four variables with, for example, the country, the site, the transect and the quadrat. A country may have several sites, and each site can include several transects, which can then include multiple quadrats. These nested variables are very important as they make it possible to integrate datasets with different spatial and taxonomic precisions. Once the variables are selected, their types (e.g. character, numeric) and units (e.g. meters) must be defined.

As previously mentioned, data can originate from four main sources: databases, data papers, research articles with associated data and unpublished raw data from data providers. The acquisition of data from the first three sources (**Fig. 2 - 1.5 to 1.7**) can be achieved through internet literature reviews and institutional or public repositories (Michener, 2018a). In addition to these approaches, the fourth source of data may necessitate a call for contribution using existing mailing lists or social networks. For this last data source, once the list of potential data providers is established, an email template is written (**Fig. 2 - 1.2**) to describe the context and the goal of the project, as well as the

181 required data (by describing the standard variables). The way that people will be cited and
182 acknowledged in the documents that will be produced by the project (e.g. publication, reports) must
183 also be addressed and should be transformed into a Data Sharing Agreement (DSA). This document
184 (see Supplementary material; **Fig. 2 - 1.3**) defines the terms of the agreement between the data
185 provider and the person or organization responsible for the project. While the DSA may not be part
186 of a legal framework, it seeks to establish a mutual agreement and terms of use for the data, building
187 the confidence and trust between the data providers and users. Emails and DSA are then sent to all
188 of the potential data providers based on the created list (**Fig. 2 - 1.4**). When received, the original raw
189 data files are then stored with the signed DSA.

190

191 2.2. Individual data reformatting

192 The next step is the individual dataset reformatting which corresponds to a standardization of the
193 variables of each dataset gathered from the four data sources. The raw data are first imported into
194 the software, either from their format of origin or by an intermediate step, where they are first
195 exported in a plain text format (e.g. csv, txt). The importation of raw data in their format of origin
196 ensures a complete reproducibility but requires the use of specific packages (e.g. “readxl” on R
197 (Wickham & Bryan, 2019)) in order to work with all of the different raw data formats. If the
198 intermediate step, where the raw data files are exported in plain text format, is chosen, all of the file
199 paths and spreadsheet names must be written to ensure the traceability of the data.

200 If the data are separated into multiple data files, usually with one main data file, and one or more
201 supplementary data files (e.g. file with sites coordinates, file with equivalence of taxonomic codes),
202 they must all be merged together (**Fig. 2 - 2.2**). Particular attention must be paid to the factor levels
203 of the grouping variable (i.e. the variable present in both files by which the merging is done) in order
204 to avoid any loss of information. Factors are a type of variable characterized by a fixed and known set
205 of possible values, that are named levels (e.g. the variable “Site” contained the levels of factors

206 “Station A” and “Station B”). A slight difference in factor levels (e.g. first letter of one in uppercase
207 and the first letter of the second in lowercase) can lead to non-matching. Multiple data files can also
208 occur when data are stored in one file but are divided into several spreadsheets. This division is
209 usually done to separate the different years or sites within the monitoring programs. Two cases can
210 be considered to address such formatting. If the data in the different spreadsheets share the exact
211 same formatting (i.e. same columns names) they can be bound together using a loop into a single,
212 long dataset. Otherwise, each spreadsheet must be treated as a different dataset.

213 Variables corresponding to the standard variables (see part 2.1 Data gathering) are first selected.
214 Then, if the data are presented in wide format (i.e. one variable divided into several columns) they
215 must be transformed into long format (i.e. one variable by column). Next, the variable corresponding
216 to the ID of the dataset is added. Here we propose a code of several letters associated with one (e.g.
217 DATA1) or more numbers if the data comes from multiple spreadsheets within a single file (e.g.
218 DATA1.1, DATA1.2, etc.). Then, all variables selected for the synthetic dataset but absent in the raw
219 data or associated metadata files, are added (**Fig. 2 - 2.3**). For example, these variables may have
220 been created specifically for the project purpose or may correspond to information given by the data
221 provider. To ensure reproducibility, we recommend that any correspondences with data providers
222 are tracked and that they are referenced by adding comments in the code. The variables are then
223 renamed to match the standard variable names. All variables containing information on the
224 taxonomic level are grouped together in a temporary variable named “Tax_ID”. The taxonomy will be
225 resolved during the following step. Next, the variables are modified to fit with the units defined. The
226 transformation involves variables such as latitude and longitude (e.g. from one coordinate reference
227 system (CRS) to another), the altitude or the depth (e.g. feet to meters), the date (e.g. DD-MM-YY to
228 YYYY-MM-DD) or the metric variable (e.g. size from mm to cm, number of individuals on the transect
229 to number of individuals on 100 m^2). When possible, we recommend that the International System of
230 Units be used. Particular attention must be given to the variable types (e.g character, numeric) of
231 each variable, as multiple data types may occur within the same variable. For example, some data

providers may have used both numeric values (e.g. “5”) and intervals (e.g. “> 5 meters”) for the depth.

Finally, once the individual data reformatting is completed, each reformatted dataset is exported in plain text format using a consistent file name nomenclature (e.g. 02_reformatted_*datasetID*).

2.3. Data grouping and taxonomic assignment

Once all data files are individually reformatted, they are all bound together (**Fig. 2 - 3.1**). This can be done automatically by retrieving all files with a specific naming structure in the storage folder and binding them using a loop. Some error messages can occur when the variable types differ between the individual reformatted files. While such errors can be avoided by converting all variables in character string, we argue that these error messages are useful to identify mistakes in variable types. If such errors happen, a modification of the code for individual datasets concerned in the previous step is necessary (**Fig. 2 - 2.3**).

As mentioned above, all information related to the taxonomy is, at this stage, stored in the unique variable (“Tax_ID”). This variable is extracted from the main data and duplicates are removed to only keep unique levels of factor (**Fig. 2 - 3.2**). Then, depending on the variables selected at the beginning of the workflow, two pathways exist: the taxonomic re-categorization (path 3A) and the taxonomic verification (path 3B).

2.3.1. Taxonomic re-categorization

The first pathway (**Fig. 2 - path 3.A**) corresponds to the case where the selected variables are not (or partially) included within the taxonomic classification system. Such cases can arise when the taxonomic identification in the field is difficult and therefore invokes the use of broad categories

instead of species or genus level ID. This is the case for coral reef benthic monitoring data with categories such as algae, rock or hard coral. In this situation, levels of factor (i.e. names of categories) are likely to vary greatly among data providers. To achieve a homogeneous classification within the synthetic dataset it is thus necessary to perform a re-categorization of the levels of factor used by each data provider. To do so, a classification is first defined by choosing standardized levels of factor, which should make it possible to re-categorize all cases. This homogeneous classification can be based on more than one variable or a set of nested variables if necessary (see part 2.1). Then, if the number of factor levels is low, the re-categorization can be done directly within the software, else a file (e.g. in csv format) containing the unique "Tax_ID" levels must be exported and the re-categorization variables must be completed manually. In the second case, the levels of factor of the "Tax_ID" variable must not be modified because they will be used to merge the re-categorization file with the main data (**Fig. 3**). The re-categorization is the most critical and the least reproducible part of the workflow and thus it must be both rigorous and consistent. Re-categorization by multiple individuals (i.e. cross-validation) would likely help to improve the reliability of this step. The following is a list of particular cases that may result for a given level of factor of "Tax_ID":

- Mixed categories. Here, the lowest common category is used. For example, if the "Tax_ID" is "Macroalgae and turf algae", the category "Algae" can be retained.
- Stacked categories. Here, the upper category can be used. For example, if the "Tax_ID" is "Algae on rock", the category "Algae" can be used.
- Homonym taxa names (e.g. "Turbinaria" which is a genus of Scleractinia but also of Fucales). In this case, it is necessary to contact the data provider and to modify the level of factor by an unambiguous one (e.g. "Algae - Turbinaria").
- Not required categories (e.g. "Shadow"). Here, the taxonomic variables remain empty, the rows which are not filled will later be removed.

Once the re-categorization is finished, the file is imported into the software and merged with the main data (**Fig. 2 - 3.6**). **Fig. 3** illustrates the taxonomic assignment for path 3.A.

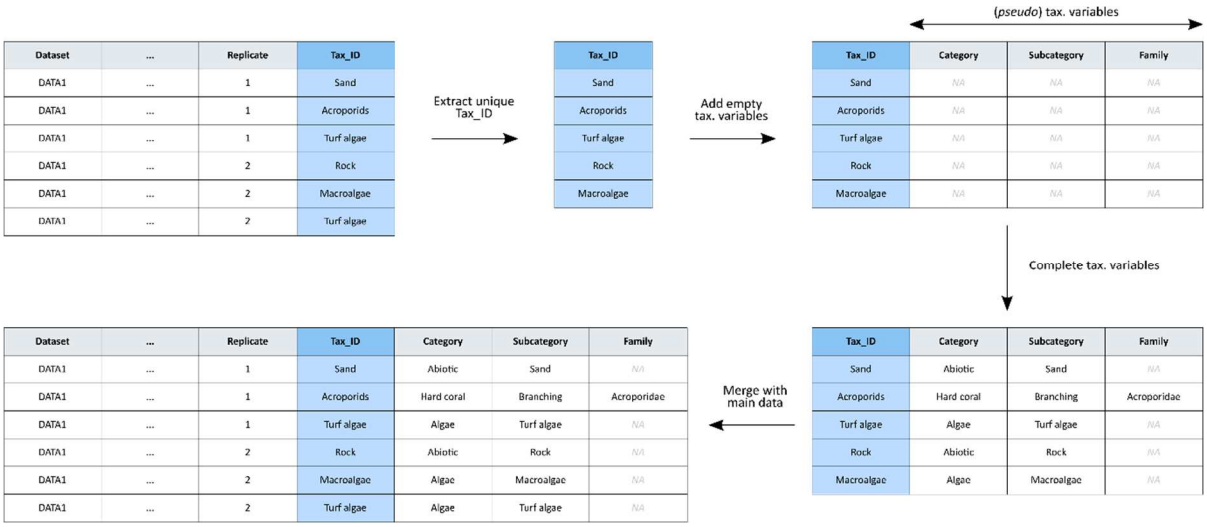


Figure 3. Example of taxonomic assignment for path 3.A (taxonomic re-categorization). NA = Not Available, tax. = taxonomic.

2.3.2. Taxonomic verification

The second pathway (**Fig. 2 - path 3.B**) corresponds to the case where the selected variables are part of the taxonomic classification system (e.g. family, genus, species). First, an API (*Application Programming Interface*, i.e. a service which allows for the query and/or upload of data from the web in a standardized format) of an online taxonomic database (e.g. “rfishbase” (Boettiger, Lang, & Wainwright, 2012) and “taxize” (Chamberlain et al., 2020) packages on R) is used to add the higher taxonomic variables for each level of the “Tax_ID” variable. The levels of factor of the “Tax_ID” variable for which upper taxonomic variables were completed through the database correspond to correct species names, while those which remained filled with NA correspond to incorrect species

names. The levels of factor with incorrect species names may be due to data entry spelling errors, changes in taxonomic classification, vernacular species names (e.g. honeycomb grouper) and incomplete (e.g. *Cephalopholis sp.*) or upper taxonomic names (e.g. Lutjanidae). Thereafter, all rows that have not been completed automatically through the API, must be manually filled. To this end, a file containing all incorrect species names is exported in a plain text format file (e.g. csv) and the higher taxonomic variables are completed for each row. When it is not possible to find the equivalence for the incorrect species name, the higher taxonomic variables are left empty (the rows which are not filled will be removed afterwards). Once finished, the file is imported into the software, bound with the table of correct species names and then merged with the main data (**Fig. 2 - 3.6**). **Fig. 4** illustrates the taxonomic assignment for path 3.B.

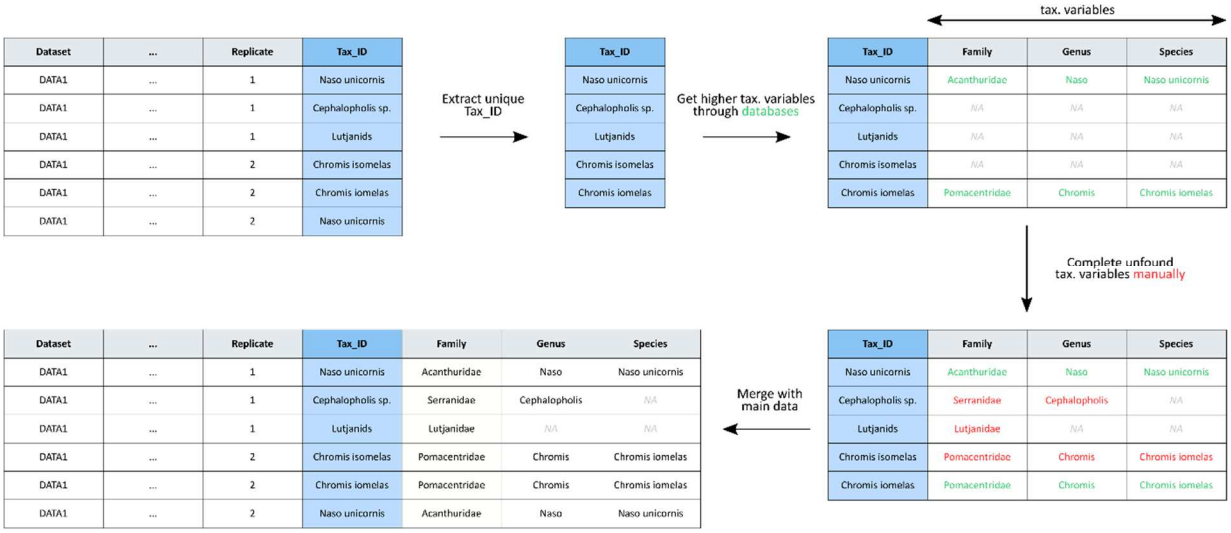


Figure 4. Example of taxonomic assignment for path 3.B (taxonomic verification). NA = Not Available, tax. = taxonomic.

Following the taxonomic assignment, it is possible that some rows for the taxonomic variables were left empty (i.e. fill with NA) because they corresponded either to not required (e.g. “Shadow”) or non-categorizable categories (e.g. “juvenile fish”). We recommend that the levels of factor of “Tax_ID” for which taxonomic variables were left empty be verified, and the corresponding rows then be deleted.

Rows that had different levels of factor for “Tax_ID” before the taxonomic assignment (e.g. “Red algae” and “Brown algae”) may have been re-categorized with a common level of factor for taxonomic variable (e.g. “Algae”) after the taxonomic assignment. Hence, it is necessary to aggregate these rows (on the metric(s) variable(s)) to avoid having multiple rows which correspond to the same category in the same sampling unit, which unnecessarily increases the size of the dataset (**Fig. 2 – 3.7**).

Finally, the “Tax_ID” variable, that was used for the taxonomic assignment but which is useless for the final synthetic dataset, is then deleted.

2.4. Quality Assurance and Quality Control

The final step of the workflow is the Quality Assurance and Quality Control (QAQC). First, duplicates are identified within the data (**Fig. 2 - 4.1**) to make sure that the same data file, or row, was not included more than once. Any duplicates found, should be removed.

Next, errors should be identified and corrected, or removed, if the error cannot be corrected (**Fig. 2 - 4.2**). Within the term “errors” we mean (1) any incorrect levels of factor for qualitative variables, (2) any incorrect values for quantitative variables and (3) any invalid site coordinates (through latitude and longitude variables). We detail these three points in the following paragraphs but more information about QAQC can be found in Michener (2018b) and Vandepitte et al. (2015).

2.4.1. Qualitative variables

The identification of incorrect levels of factor can be done visually by checking the list of unique levels of each qualitative variable. If the number of levels of factor is too high for a full visual inspection, a list of selected levels can be created for each qualitative variable and compared with the actual levels of factor present in each. Once identified, levels must be corrected during the individual data reformatting step (**Fig. 2 - 2.3**).

2.4.2. Quantitative variables

For quantitative variables, we distinguish incorrect values which are values that are outside of the normal range (e.g. a negative abundance) from outliers which are extreme values (e.g. a value of 100 while the mean is 1). Incorrect values are easily identifiable for variables whose values fall within an interval of one (e.g. [0 ; +∞]) or two limits (e.g. [0 ; 100]). A process can be put in place to automatically remove incorrect values (i.e. those below or above the limit(s)) but we recommend that each value be checked individually in order to correct for any errors that may have resulted from the individual reformatting step. Outliers, however, are more difficult to treat. While their identification through graphical (e.g. boxplot) or statistical (e.g. Grubbs' test, Dixon's Q-test) methods is easy, it is generally difficult to know if their values result from natural variability or from an error made during or following the acquisition of the data. Thus, these outliers must not be arbitrarily removed from the data, without evidence which proves that they actually resulted from an error. To this end, we recommend that the data provider be contacted in order to benefit from his/her expertise. Among the errors which can be committed during the individual data reformatting step (**Fig. 2 - 2.3**), unit conversion mistakes are the most common. Particular attention must be given to the metric variable(s) (i.e. response variable) and the QAQC method must be specifically developed for each of them. For example, for data expressed in percentage of cover, a quadrat can be divided

into multiple rows where each corresponds to the cover of a given taxa. Hence, it is necessary to aggregate the cover percentage of these rows to verify that the total cover of the quadrat remains above 0 % and below 100 %. We highly recommend the use of interactive data visualisation tools, such as HTML tables (e.g. “DT” on R (Xie, Cheng, & Tan, 2020)) or interactive plots (e.g. “plotly” on R (Sievert, 2020)), to help identify incorrect values (see Supplementary material).

2.4.3. Site coordinates

Errors in data entry or an incorrect transformation of coordinates (e.g. non-homogenized coordinate reference systems) may cause the presence of invalid site coordinates. The identification of some invalid positions can be assessed through visual inspection, such as by using an interactive map (e.g. R package “leaflet”, (Cheng, Karambelkar, & Xie, 2019)). However, the response time of interactive maps tends to increase with the number of site coordinates. Thus, an alternative is to perform this verification for each individual dataset (**Fig. 2 - 2.3**), when the number of coordinates is still manageable. The advantage of this method is that it allows for site positions to be confirmed to ensure that they are consistent within the geographic extent of each dataset. While some invalid site coordinates are easily identifiable (e.g. located on land while the data come from marine monitoring programs) others can be more difficult to identify, for example when the error in latitude and longitude is so small that it leads to a slight modification of the site position. Finally, a broad automatic identification can be done using a polygon shapefile representing the area in which the sites are supposed to be present, at which point sites that do not fall inside this polygon can then be identified. Unfortunately, this method cannot totally replace a visual inspection. Once identified, the correction of invalid site coordinates must be done during the individual data reformatting step (**Fig. 2 - 2.3**).

In addition to the QAQC step, we strongly encourage the data aggregator to send basic data visualisation and identified errors of individually integrated datasets to data providers and to ask for their feedback in order to ensure accuracy.

2.4.4. Export of the synthetic dataset

Finally, when all of the errors are corrected, the final synthetic dataset is exported (**Fig. 2 - 4.3**). We strongly recommend associating a metadata file to the synthetic dataset, where at least a description of the variables and their units are provided. However, if the synthetic dataset is to be shared, a more complete description of the data is required and the use of the Ecological Metadata Language (EML) can be considered for this purpose (Fegraus, Andelman, Jones, & Schildhauer, 2005). Metadata can include the name and contact of data aggregator(s), the link for the code repository used for data integration, a full description of individual datasets integrated, as well an appropriate reference to cite the synthetic dataset.

3. Case studies

To illustrate our workflow and to facilitate its use, we have provided an R code template for two case studies. These case studies are inspired by the *Status of Coral Reefs of the World: 2020* report for which 248 datasets of coral reef benthic monitoring data were integrated to assess the global status and trends of hard coral cover over a period of more than three decades (Souter et al., 2021). However, as the data used in the frame of this report were gathered through DSA and were

restricted in their use, we created example datasets that do not correspond to real datasets but illustrate the main types of data formats encountered in this project. The first case study illustrates the integration of data from the monitoring of benthic communities (sessile organisms) in coral reefs. It corresponds to path 3A of the workflow, with a taxonomic re-categorization. The second case study illustrates the integration of data from the monitoring of coral reef fish communities (mobile organisms). It corresponds to path 3B of the workflow, with a taxonomic verification. The code template provided illustrates steps 2 (individual data reformatting), 3 (data grouping and taxonomic assignment) and 4 (QAQC) of the workflow, as the first step doesn't involve code. Each of these three steps are associated with an R script. The *.Rmd* format (rmarkdown (Xie, Allaire, & Grolemund, 2018)) was selected for the different R scripts because it allows for a better segmentation and annotation of the code and process (necessary for the second step) and for the exportation of code and output to an HTML file which may include interactive tables, plots and maps (necessary for the last two steps, see Supplementary material). The code template for these two case studies and associated information are available at https://github.com/JWicquart/monitoring_workflow.

4. Discussion

4.1. Lessons learned and limits of the workflow

The workflow design was possible thanks to the development of recent packages in R which make data wrangling easier, and which facilitate the access to online taxonomic databases and promote interactive data visualisation. Despite these technical improvements, data integration remains a time consuming task, as some steps are difficult to automate, and a full reproducibility can sometimes be hard to achieve.

Within the context of the *Status of Coral Reefs of the World: 2020* report (Souter et al., 2021), the vast majority of the 248 datasets that were integrated were unpublished data which came from

numerous data providers. For this reason, over the entire year that was necessary to one person to complete the data integration, the data gathering has represented the longest step in the workflow, involving the identification and contact of potential contributors, signing DSAs, and finally engaging in discussions with contributors to ensure that their data was properly reformatted. Despite their financial cost, the organization of several workshops in different countries increased the visibility of the project, which significantly increased the number of data providers and facilitated the individual data reformatting step. However, in contrast to data publication, data sharing, on which the *Status of Coral Reefs of the World: 2020* report (Souter et al., 2021) was mainly built, limits the traceability of integrated individual datasets (as they are not associated with a DOI), reducing the reproducibility of the workflow. The trend towards increased data publication should nevertheless help to reduce this problem in the coming years (Costello et al., 2013; Shin et al., 2020).

The reformatting of individual data has also constituted a time-consuming step, requiring between thirty minutes to several hours per dataset, depending on the complexity and machine readability of the data format as well as the completeness of the metadata provided. Lack of data standards (e.g. Darwin Core) and appropriate data management practices are thus the main factors which explain the time required for this step and the difficulty for a complete automation. With regards to the reproducibility of this step, it can be improved by describing decisions taken to correct errors for each individual dataset, either directly in the code, or in metadata associated with the synthetic dataset.

Given the difficulty of identifying benthic organisms on coral reefs, each monitoring program defined and used its own nomenclature, using broad benthic categories instead of the taxonomic classification system. For that reason, in the context of the *Status of Coral Reefs of the World: 2020* report (Souter et al., 2021), and for the third step of the workflow (path 3.A), we defined a nested classification which allowed for the re-categorization all the broad benthic categories used in the individual datasets. This process required an ecological expertise and was iterative, as the chosen

nested classification was updated several times, when new cases were encountered. Once the nested classification was defined, a manual stage was necessary to re-categorize categories used in the individual datasets into the categories selected. This operation can be more or less laborious depending on the number or categories to re-categorize. However, the taxonomic assignment step can be quick and almost fully automated in cases where taxonomic verification is used (path 3.B, true taxonomic categories) instead of taxonomic re-categorisation (path 3.A, non-taxonomic categories).

Finally, for the QAQC step, as the data integration process used for the *Status of Coral Reefs of the World: 2020* report (Souter et al., 2021) aimed to provide a ready-to-use synthetic dataset for data analysis, we deleted all data that could not have been corrected, which sometimes led to a loss up to 10% of an individual dataset. Our first case study can be easily adapted for the data integration of sessile communities (with metrics such as percentage cover) in the marine and terrestrial realms. However, we caution that for mobile communities (with metrics such as abundance or size), like fish or birds, the QAQC process is more difficult to perform as the data quality depends heavily on the sampling strategy. Nonetheless, it is important to note that all issues raised by the data heterogeneity cannot be fully resolved during the data integration process but that some of them may be addressed during the analysis itself. Analytical methods of synthetic datasets are beyond the scope of this article and information on this subject can be found in Recknagel and Michener (2018). Furthermore, different analytical methods can be used to limit bias relative to data heterogeneity, such as the sensitivity analysis which allows for the identification of datasets which have a greater influence on observed trends.

4.2. Advantages and comparison with existing approaches

In spite of the limitations mentioned, which mainly concern the time-consuming nature of the process, this workflow presents several advantages. First, because it is founded on the direct

acquisition of the data from people in charge of their collection, it widens the scope for data that can be integrated to those that have not been published and enhances the quality of the data integration by benefiting from the expertise of data providers. Second, the workflow allows for the integration of different levels of precision of taxonomic and spatial data, elements that generally vary between monitoring programs. Because this method allows for raw data to be integrated, changes can be directly assessed in the unit of the considered metric (i.e. full-data analysis (Spake et al., 2020)) instead of using effect-size, which is usually used in ecological syntheses through meta-analyses (e.g. Côté, Gill, Gardner, & Watkinson (2005)). Finally, this method guarantees a high level of reproducibility in facilitating the identification and correction of errors committed during the data integration process. Workflows are particularly relevant for reproducibility (Poisot et al., 2016; Cohen-Boulakia et al., 2017; Botvinik-Nezer et al., 2020) as they provide a visual representation of the different steps, and the possibility to adapt each step depending on the goals of the project (Jones et al., 2006).

Until now, a vast majority of bottom-up approaches for data integration were developed by large databases, in particular GBIF (GBIF: The Global Biodiversity Information Facility, 2021) or OBIS (OBIS, 2021). These databases typically include a web-based interface to allow data providers to publish their datasets, such as the Integrated Publishing Toolkit (IPT) developed by GBIF (Robertson et al., 2014), and are based on data standards, such as the DarwinCore (Wieczorek et al., 2012). Unlike these databases, which seek to increase the accessibility and standardization of biodiversity datasets, the workflow presented here aims to produce a ready-to-use synthetic dataset, adapted to the analyses that will be performed, but also to integrate datasets whose formats are not yet suited for incorporation into existing databases.

Outside of large databases, the number of studies related to data integration remains limited and have been mainly focused on quality control (e.g. (Dou et al., 2012; Belitz et al., 2018)). Most have used similar steps to the one presented here to control data quality, and to check geography,

taxonomy and the completeness of data or to investigate outliers (Dou et al., 2012; Vandepitte et al., 2015). These studies also tended to explicitly state quality checks in the form of a series of questions or items to review (Vandepitte et al., 2015; O'Donnell et al., 2021). As these items must be specifically tailored to the data and the analyses that follow, we did not explicitly present a list of quality checks here, but we strongly recommend the implementation of this approach. Several studies have highlighted the requirement of manual step(s) during quality control, where the knowledge of the data aggregator is essential to verify the consistency and ecological relevance of modifications applied (Jones et al., 2006). Overall, the individual data reformatting step is not mentioned in studies related to data integration mainly because the approaches developed are built on web-based interfaces (Chaudhary et al., 2010; Robertson et al., 2014) where the data providers are left with the responsibility of reformatting their data. Finally, O'Donnell et al., (2021) presented a similar approach to the one presented here, where they developed a framework and a custom open-source software to integrate long-term monitoring data of the greater sage-grouse population. While this framework and resulting considerations could be applied to other studies (O'Donnell et al., 2021), the software they developed was tailored to their specific scenario and could be difficult to reuse. Based on the R code associated with the two case studies, we hope that the workflow presented here can bridge this gap, through a more versatile approach.

5. Conclusion

The purpose of the workflow presented in this study is to encourage researchers to integrate multiple ecological monitoring data into a single synthetic dataset in order to perform analyses on the status and trends in biodiversity and ecosystems at larger scales. The results of these analyses are essential to inform policy makers and to measure the effectiveness of global (e.g. CBD Post-2020 Biodiversity Framework), regional or national programs, that aim to protect biodiversity and ecosystems (Balmford, Green, & Jenkins, 2003). Data integration could also be of great interest for

research topics, such as macroecology and biogeography, by testing ecological hypotheses at larger scales (Carpenter et al., 2009; Poisot et al., 2016; König et al., 2019). The need for data integration will likely increase in the coming years (Miller et al., 2019), due to the emergence of new standards on data interoperability (e.g. FAIR data principles, (Wilkinson et al., 2016)) and a trend towards further data sharing (Michener, 2015) or data publication (Costello et al., 2013; Shin et al., 2020), which together sharply increase the volume of accessible data available to the scientific community (Hampton et al., 2013). In this context, the publication of code and workflow are essential to increase the reproducibility of results and to drive ecology toward a more transparent science.

Acknowledgments

We thank the International Coral Reef Initiative (ICRI) and the Prince Albert II of Monaco Foundation for the funding. We thank Claire Bissery, Charlotte Moritz, Jason Vii, Mary Donovan and all the people involved in the *Status of Coral Reefs of the World: 2020* report for the valuable discussions we had during the design of the workflow. We thank two anonymous reviewers for their helpful comments.

Authors' contributions

JW designed the workflow, made the figures and wrote the code template for the two case studies. JW and SP led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

References

Balmford, A., Bennun, L., Ten Brink, B., Cooper, D., Côté, I. M., Crane, P., … Walther, B. A. (2005). The convention on biological diversity's 2010 target. *Science*, *307*(5707), 212–213. doi:10.1126/science.1106281

Balmford, A., Green, R. E., & Jenkins, M. (2003). Measuring the changing state of nature. *Trends in Ecology and Evolution*, *18*(7), 326–330. doi:10.1016/S0169-5347(03)00067-3

Belitz, M. W., Hendrick, L. K., Monfils, M. J., Cuthrell, D. L., Marshall, C. J., Kawahara, A. Y., … Monfils, A. K. (2018). Aggregated occurrence records of the federally endangered Poweshiek skipperling (Oarisma poweshiek). *Biodiversity Data Journal*, *6*. doi:10.3897/BDJ.6.e29081

Boettiger, C., Lang, D. T., & Wainwright, P. C. (2012). rfishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, *81*(6), 2030–2039. doi:10.1111/j.1095-8649.2012.03464.x

Borregaard, M. K., & Hart, E. M. (2016). Towards a more reproducible ecology. *Ecography*, *39*, 349–353. doi:10.1111/ecog.02493

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Ball, S. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. doi:10.1038/s41586-020-2314-9

Carpenter, S. R., Armbrust, E. V., Arzberger, P. W., Chapin, F. S., Elser, J. J., Hackett, E. J., … Zimmerman, A. S. (2009). Accelerate Synthesis in Ecology and Environmental Sciences. *BioScience*, *59*(8), 699–701. doi:10.1525/bio.2009.59.8.11

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., … Grenié, M. (2020). taxize: Taxonomic information from around the web. Retrieved from https://github.com/ropensci/taxize

Chaudhary, B. V., Walters, L. L., Bever, J. D., Hoeksema, J. D., & Wilson, G. W. T. (2010). Advancing synthetic ecology: a database system to facilitate complex ecological meta-analyses. *The Bulletin of the Ecological Society of America*, *91*(2), 235–243. doi:10.1002/bes2.1214

Cheng, J., Karambelkar, B., & Xie, Y. (2019). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. Retrieved from https://cran.r-project.org/package=leaflet

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., … Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. doi:10.1016/j.future.2017.01.012

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, *28*(8), 454–461. doi:10.1016/j.tree.2013.05.002

Côté, I. M., Gill, J. A., Gardner, T. A., & Watkinson, A. R. (2005). Measuring coral reef decline through meta-analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 385–395. doi:10.1098/rstb.2004.1591

Dornelas, M., Antão, L. H., Moyes, F., Bates, A. E., Magurran, A. E., Adam, D., … Zettler, M. L. (2018). BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography*, *27*(7), 760–786. doi:10.1111/geb.12729

Dou, L., Cao, G., Morris, P. J., Morris, R. A., Ludäscher, B., Macklin, J. A., & Hanken, J. (2012). Kurator: A Kepler package for data curation workflows. *Procedia Computer Science*, *9*, 1614–1619. doi:10.1016/j.procs.2012.04.177

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to ecological metadata language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, *86*(3), 158–168.

GBIF: The Global Biodiversity Information Facility. (2021). What is GBIF?. Available from https://www.gbif.org/what-is-gbif [27 September 2021].

Hallmann, C. A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H., … de Kroon, H. (2017). More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE*, *12*(10).

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., … Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. doi:10.1890/120103

Henry, P.-Y., Lengyel, S., Nowicki, P., Julliard, R., Clobert, J., Čelik, T., … Henle, K. (2008). Integrating ongoing biodiversity monitoring: Potential benefits and methods. *Biodiversity and Conservation*, *17*(14), 3357–3382. doi:10.1007/s10531-008-9417-1

Jones, M. B., Schildhauer, M. P., Reichman, O. J., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, *37*(2006), 519–544. doi:10.1146/annurev.ecolsys.37.091305.110031

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration - the significance of data resolution and domain. *PLoS Biology*, 1–16.

Külh, H. S., Bowler, D. E., Bösh, L., Bruelheide, H., Dauber, J., & Eichenberg, D. (2020). Effective Biodiversity Monitoring Needs a Culture of Integration. *One Earth*, *3*(4), 462–474. doi:10.1016/j.oneear.2020.09.010

Lindenmayer, D. B., & Likens, G. E. (2009). Adaptive monitoring: a new paradigm for long-term research and monitoring. *Trends in Ecology and Evolution*, *24*(9), 482–486. doi:10.1016/j.tree.2009.03.005

Lindenmayer, D. B., & Likens, G. E. (2010). The science and application of ecological monitoring. *Biological Conservation*, *143*(6), 1317–1328. doi:10.1016/j.biocon.2010.02.013

Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44. doi:10.1016/j.ecoinf.2015.06.010

Michener, W. K. (2018a). Data discovery. In F. Recknagel & W. K. Michener (Eds.), *Ecological Informatics* (pp. 115–128). Springer.

Michener, W. K. (2018b). Quality assurance and quality control (QA/QC). In F. Recknagel & W. K. Michener (Eds.), *Ecological Informatics* (pp. 55–70). Springer.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, *27*(2), 85–93. doi:10.1016/j.tree.2011.11.016

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. doi:10.1111/2041-210X.13110

O'Donnell, M. S., Edmunds, D. R., Aldridge, C. L., Heinrichs, J. A., Monroe, A. P., Coates, P. S., … Wightman, C. S. (2021). Synthesizing and analyzing long-term monitoring data: A greater sage-grouse case study. *Ecological Informatics*, *63*, 101327. doi:10.1016/j.ecoinf.2021.101327

OBIS. (2021). Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. www.obis.org.

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., & Peres-Neto, P. (2019). Ecological Data Should Not Be So Hard to Find and Reuse. *Trends in Ecology and Evolution*, *34*(6), 494–496. doi:10.1016/j.tree.2019.04.005

Poisot, T., Gravel, D., Leroux, S., Wood, S. A., Fortin, M., Baiser, B., … Stouffer, D. B. (2016). Synthetic datasets and community tools for the rapid testing of ecological hypotheses. *Ecography*, *39*(4), 402–408. doi:10.1111/ecog.01941

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Recknagel, F., & Michener, W. K. (Eds.). (2018). *Ecological Informatics: Data Management and Knowledge Discovery* (Third Edit). Springer. doi:10.1016/b978-008045405-4.00170-1

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, *331*(February), 703–706.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., … Desmet, P. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLoS ONE*, *9*(8). doi:10.1371/journal.pone.0102623

Schildhauer, M. (2018). Data integration: Principles and practice. In F. Recknagel & W. K. Michener (Eds.), *Ecological Informatics* (pp. 129–157). Springer.

Schmeller, D. S., Julliard, R., Bellingham, P. J., Böhm, M., Brummitt, N., Chiarucci, A., … Belnap, J. (2015). Towards a global terrestrial species monitoring program. *Journal for Nature Conservation*, *25*, 51–57. doi:10.1016/j.jnc.2015.03.003

Shin, N., Shibata, H., Osawa, T., Yamakita, T., Nakamura, M., & Kenta, T. (2020). Toward more data publication of long-term ecological observations. *Ecological Research*, *35*(5), 700–707. doi:10.1111/1440-1703.12115

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. Retrieved from https://plotly-r.com

Souter, D., Planes, S., Wicquart, J., Logan, M., Obura, D., & Staub, F. (2021). *Status of Coral Reefs of the World: 2020*. Global Coral Reef Monitoring Network.

Spake, R., Mori, A. S., Beckmann, M., Martin, P. A., Christie, A. P., Duguid, M. C., & Doncaster, P. C. (2020). Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecology Letters*, *24*(2), 374–390. doi:10.1111/ele.13641

Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., … Mees, J. (2015). Fishing for data and sorting the catch: Assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. *Database*, *2015*. doi:10.1093/database/bau125

Vanderbilt, K., & Gaiser, E. (2017). The International Long Term Ecological Research Network: A platform for collaboration. *Ecosphere*, *8*(2). doi:10.1002/ecs2.1697

Vos, P., Meelis, E., & Ter Keurs, W. J. (2000). A framework for the design of ecological monitoring

677 programs as a tool for environmental and nature management. *Environmental Monitoring and*
678 *Assessment*, *61*(3), 317–344. doi:10.1023/A:1006139412372

679 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019).
680 Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.
681 doi:10.21105/joss.01686

682 Wickham, H., & Bryan, J. (2019). readxl: Read Excel Files. Retrieved from https://cran.r-
683 project.org/package=readxl

684 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., … Vieglais, D. (2012).
685 Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, *7*(1).
686 doi:10.1371/journal.pone.0029715

687 Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Gabrielle, A., Axton, M., Baak, A., … Mons, B.
688 (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific*
689 *Data*, *3*, 1–9. doi:10.1038/sdata.2016.18

690 Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Boca Raton, Florida:
691 Chapman and Hall/CRC. Retrieved from https://bookdown.org/yihui/rmarkdown

692 Xie, Y., Cheng, J., & Tan, X. (2020). DT: A Wrapper of the JavaScript Library 'DataTables'. Retrieved
693 from https://cran.r-project.org/package=DT

694