

# popler: An R package for synthesis of population time series from long-term ecological research

Aldo Compagnoni<sup>\*a,b,c</sup>, Andrew J. Bibian<sup>a</sup>, Brad M. Ochocki<sup>a</sup>, Sam Levin<sup>b,c</sup>, Kai  
Zhu<sup>d</sup> and Tom E.X. Miller<sup>a</sup>

<sup>a</sup>Department of BioSciences, Program in Ecology and Evolutionary Biology, Rice  
University, 6100 Main St, MS-170, Houston, TX 77005

<sup>b</sup>Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor  
1, 06108 Halle (Saale), Germany

<sup>c</sup>German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,  
Deutscher Platz 5e, 04103 Leipzig, Germany

<sup>d</sup>Department of Environmental Studies, University of California, Santa Cruz, CA  
95064, USA

Running headline: The popler database and R package

---

<sup>\*</sup>aldo.compagnoni@idiv.de

## Abstract

1. Population dynamics play a central role in the historical and current development of fundamental and applied ecological science. The nascent culture of open data promises to increase the value of population dynamics studies to the field of ecology. However, synthesis of population data is constrained by the difficulty in identifying relevant datasets, by the heterogeneity of available data, and by access to raw (as opposed to aggregated or derived) observations.
2. To obviate these issues, we built a relational database, `popler`, and its R client, `library("popler")`. `popler` accommodates the vast majority of population data under a common structure, and without the need for aggregating raw observations. `library("popler")` is designed for users unfamiliar with the structure of the database and with the SQL language. This R library allows users to identify, download, explore, and cite datasets salient to their needs.
3. We implemented `popler` as a PostgreSQL instance, where we stored population data originated by the United States Long Term Ecological Research (LTER) Network. Our focus on the US LTER data aims to leverage the untapped potential of this vast open data resource. The database currently contains 306 datasets from 25 LTER sites. `popler` is designed to accommodate automatic updates of existing datasets, and to accommodate additional datasets from LTER as well as non-LTER studies.
4. The combination of the online database and the R library `library("popler")` is a resource for data synthesis efforts in population ecology. The common structure of `popler` simplifies comparative analyses, and the availability of raw data confers flexibility in data analysis. `library("popler")` maximizes these opportunities by providing a user-friendly interface to the online database.

## Keywords

- 1 open long-term population data, US Long Term Ecological Research Network data, online
- 2 database, database structure, PostgreSQL, R package, data synthesis, comparative analysis

### 3 Introduction

Population dynamics – changes in species’ abundance and composition through time and space – are central to ecology for both applied and fundamental reasons. Populations are the building blocks of ecological dynamics at higher scales of organization, and examples abound showing how the study of population ecology improves understanding in evolution (Metcalf and Pavard, 2007), community ecology (Levine and HilleRisLambers, 2009), and ecosystem ecology (Medvigy et al., 2009; Fisher et al., 2018). Given their central role, studies of population dynamics will be an essential component in the advances allowed by the flourishing culture of open access and data synthesis.

The increase in freely available data is poised to change ecological science (Laurance et al., 2016). The rising focus on open data is clear in changing publishing standards, in the design of observational networks (Schimel et al., 2007), and in the availability of previously proprietary data (Kratz et al., 2003; Bechtold et al., 2005). This deluge of open data holds promise to facilitate comparative analyses and to test the generality of ecological hypotheses. For population dynamics in particular, it is the increasing availability of long-term data that will likely yield the most substantial scientific advances, as long time series are required to detect trends in abundance (Lindenmayer et al., 2012), quantify temporal variance (Compagnoni et al., 2016), and identify endogenous (Knape and de Valpine, 2012) or exogenous (Hampton et al., 2013) drivers of population fluctuations.

There are currently three public databases that provide time series of population data. These are the Global Population Dynamics Database (GPDD, Inchausti and Halley, 2001), the Living Planet Index (Loh et al., 2005), and BioTIME (Dornelas et al., 2018). These databases are an important resource for population biologists (e.g., Knape and de Valpine, 2012), but their characteristics make them optimal for a specific set of analyses. For example, GPDD time series refer to vertebrate species only, and there is one observation of population size or density for each temporal replicate. Alternatively, the time series in BioTIME have only yearly temporal resolution, and they are limited to two levels of nested spatial replication.

One of the best sources of publicly available long-term data is the Long-Term Ecological Research (LTER) network. The LTER was founded in 1980 and grew from the original six sites

32 to the current 28 sites throughout North America plus one each in Puerto Rico and Antarctica.  
33 Synthetic and comparative studies from the LTER network have made valuable contributions to  
34 ecological understanding (Knapp et al., 2012). However, the majority of LTER synthesis research  
35 has focused on ecological dynamics at the community (e.g. Wilcox et al. (2017)) and ecosys-  
36 tem (e.g. Knapp and Smith (2001)) scales. Nevertheless, every LTER site collects population  
37 abundance data as one of its five core areas of continuous observations (Callahan, 1984). These  
38 population time series include both single- and multi-species studies. In our opinion these data,  
39 which have been accumulating since 1980, are under-used.

40 One issue that may limit the use of LTER population data in synthetic, comparative studies  
41 is their heterogeneity. The structure of LTER data sets may be widely different, employing a  
42 variety of data types (counts of individuals, biomass estimates, percent cover, etc.), experimental  
43 designs driven by the priorities of particular PIs, and diverse replication schemes – idiosyncrasies  
44 that may be difficult to accommodate in a one-size-fits-all database. However, these challenges  
45 also present valuable opportunities. For example, the hierarchical replication structure of many  
46 LTER studies (e.g., subplots within plots within transects) can facilitate more sophisticated  
47 statistical investigation than would be possible with simpler, aggregated, or unreplicated data.

48 To overcome the issues posed by heterogeneous data structures, we developed `popler` (POP-  
49 ulation dynamics in Long-term Ecological Research), an online database of LTER population  
50 studies. This database defines a common data structure that can accommodate in principle all  
51 population data, and its SQL environment allows updates whenever new data becomes available.  
52 We also developed a companion R package to facilitate the identification, access, and manipula-  
53 tion of raw and heterogeneous population data. Our goals here are to provide introductions to  
54 the database and package. We focus on LTER time series, but expanding `popler` beyond the  
55 LTER network is a priority for future development.

## 56 The `popler` database

57 To combine population data from the LTER network using a common structure, we identified  
58 a set of relevant variables (Table 1) and organized them into a relational database (Fig. 1).  
59 We store “raw” data, meaning that we have not modified, edited, or aggregated the original

60 observations.

61 For inclusion in `popler`, we only considered studies that included (1) repeated observations  
62 of populations or individuals through time, (2) at least five population censuses (as of database  
63 creation in 2017), and (3) taxonomic information associated with abundance observations (e.g.,  
64 we excluded time series of functional groups). We provide technical details of database creation  
65 in Appendix S1.

66 The `popler` database currently contains data from 306 studies (122 of which are experi-  
67 mental) representing 4386 cumulative years of observations with a mean study duration of 14.33  
68 years. `popler` contains data from 676 plant species, 333 animal species, and 1 fungal species.

## 69 Population data

70 We define “population data” as time-series of observations on the size or density of a population  
71 of a species or other taxonomic unit. Observations of population size are stored in a variable  
72 called `abundance_observation` and can be measured as a count, biomass, density, or cover.  
73 These four types of population data are stored in the homonymous tables of the database (Fig.  
74 1A).

75 The population datasets contained in `popler` are always replicated temporally. Temporal  
76 replicates are identified with up to three variables: `year`, `month`, and `day`. Population data are  
77 also almost always spatially replicated, and spatial replicates are often nested, where for example  
78 a study might include separate sites, each of which contains intermediate spatial replicates (e.g.  
79 a transect, a block), which in turn contain the smallest spatial replicate at which observations are  
80 made (e.g. a plot, a quadrat). The hypothetical study described above would have three nested  
81 levels of spatial replication, identified by three numbered `spatial_replication` variables.  
82 In `popler`, we accommodate data sets with up to five spatial replication levels (Table 1). We  
83 call the first and therefore largest spatial replicate “study site” (Fig. 1C). Note that this does  
84 not refer to the LTER site, one of the 28 NSF-supported locations (Table S3).

85 `popler` contains both observational and experimental studies. Experimental datasets con-  
86 tain information on one or more experimental treatments. `Popler` accommodates information on  
87 up to three experimental treatments, identified by three numbered `treatment_type` variables

88 (Table 1).

89 Most datasets also contain one or more variables in addition to the ones described above  
90 which we store in a list of variables called `covariates`. Covariates can be useful for time series  
91 that contain information on population structure. In these datasets, observations on population  
92 size are grouped based on subdivisions of the entire population, such as males and females, large  
93 and small individuals, etc. We identify these datasets through the variable `structured_data`.

94 Finally, in addition to time series of abundance, `popler` contains individual-level data. This  
95 data provides information on the attributes of the individuals, or a subset thereof, that make up a  
96 population. We store this information in a dedicated table ("Individual", Fig. 1A). As individual  
97 attributes we consider variables that describe identity, size, sex, life stage or status (e.g. repro-  
98 ductive or non-reproductive). We refer to these individual attributes with the term "structure":  
99 `popler` accommodates data sets that measure up to four types of structure simultaneously. We  
100 store these data in up to four numbered `structure_type` variables. While these data are not  
101 population time series, we chose to include them in `popler` because they provide information on  
102 demographic transitions that can be used to derive estimates of population growth. Moreover,  
103 in the cases of datasets that sample all of the individuals in a population, individuals can be  
104 aggregated (i.e. summed) as a measure of population size.

## 105 **Taxonomic information**

106 Each observation corresponds to a taxonomic unit (Fig. 1B), typically a species or a genus,  
107 but also include data that refer to a higher taxonomic rank, such as family, or order. `popler`  
108 provides 15 taxonomic ranks, and two additional variables that refer to how taxonomic infor-  
109 mation is recorded in the original datasets. The additional variables are `sppcode`, which are  
110 taxon-specific alphanumeric codes, and `common_name`, the common name of each taxonomic  
111 unit (Table S1). `popler` also allows to store accepted taxonomic information in an additional  
112 table (Fig. 1B). This table accounts for ambiguities contained in the raw taxonomic data, which  
113 originate by the dynamic changes in species classifications (Chamberlain and Szöcs, 2013). Fur-  
114 ther versions of `popler` will populate this second table with the accepted taxonomic units (which  
115 include taxonomic information above the level of genus) provided by the R package `taxize`

116 (Chamberlain and Szöcs, 2013).

## 117 **Study site**

118 We stored the locations of datasets by recording the latitude (`lat_study_site`) and longitude  
119 (`lng_study_site`) of study sites (Fig. 1C). Storing this information in a separate table allows  
120 for explicit connections between independent data sets collected at the same locations within  
121 LTER sites.

## 122 **Metadata**

123 The metadata table (Table S2) provides information on temporal and spatial replication, and  
124 study design (Fig. 1D), including title, link to online metadata, contact information for data  
125 originators, and the type of data provided by the dataset (i.e., which of the five tables in  
126 Fig. 1A the data is stored in). All remaining metadata is related to the variables stored in  
127 the tables of 1A and 1B. First, we provide the years elapsed between the first and last ob-  
128 servation (`duration_years`), and the sampling frequency (`samplefreq`). Second, we pro-  
129 vide the number of levels of nested spatial replicates, and the number of replicates for each  
130 spatially nested level. Third, we show whether studies focus on a single species or on mul-  
131 tiple species through the `community` variable. Fourth, we identify studies as observational  
132 or experimental (`studytype`). If a study is experimental, we provide information on the  
133 type of treatments imposed by the study (`treatment_type_n`) and, when available, which  
134 one is the control treatment (`control_group`). Finally, when abundance data stored in the  
135 `abundance_observation` variable is aggregated across space or time, rather than raw, we  
136 consider these data as derived (`derived`).

## 137 **The `popler` package**

138 The `popler` R package consists of three core functions that allow users to browse and retrieve  
139 data from the database (Fig. 2). In order of intended use, these functions are: `pplr_dictionary()`,  
140 `pplr_browse()`, and `pplr_get_data()`.



## 141 The `pplr_dictionary()` function

142 The dictionary function is a good place for new users to begin working with `popler` (Fig.  
143 2). With no arguments provided, this function returns a subset of the most useful metadata  
144 variables associated with each dataset (Fig. 1). Providing argument `full_tbl = TRUE` returns  
145 all 77 metadata variables. Each one of these variable names can be provided as an argument  
146 to `pplr_dictionary()`, which then returns the possible unique values of the variable. For  
147 example, `pplr_dictionary(lterid)` returns the three letter codes of the LTER network sites  
148 included in `popler`. For numeric variables such as `duration_years`, `pplr_dictionary()`  
149 returns a summary including quantiles, mean, and median.

## 150 The `pplr_browse()` function

151 Once the user is familiar with the meaning and content of the variables that define `popler`  
152 datasets, they are ready to dig deeper using `pplr_browse()` (Fig. 2). Running `pplr_browse()`  
153 without arguments provides the metadata from the entire contents of the database. This will be  
154 a  $306 \times 20$  data frame, with each row corresponding to a study and each column corresponding  
155 to a variable defined by `pplr_dictionary()`.

156 The full strength of `pplr_browse()` is achieved by subsetting studies according to desired  
157 criteria using logical expressions. For example, the user might want to consider only studies  
158 whose duration is 30 years or greater, which can be subsetted with:

```
LTER_30 <- ppplr_browse( duration_years > 29)
```

159 This operation will create the object `LTER_30`, which provides metadata for the data sets  
160 that satisfy the specified criterion. Multiple criteria may be combined. For example, 30+ year  
161 studies of plants can be browsed with

```
LTER_30_plants <- ppplr_browse( duration_years > 29 &  
                                kingdom == "Plantae")
```

162 To facilitate data exploration, `pplr_browse()` output can be printed in a more readable  
163 settings by providing `report = TRUE` as an argument, which opens up a formatted html doc-  
164 ument. The metadata provided by `pplr_browse()` not only contains information on the

165 characteristics of a study but also information on how to cite the study, its unique identifiers,  
166 including digital object identifier (DOI), and the contact information of study PIs.

## 167 **The `pplr_get_data()` function**

168 Once data sets of interest have been identified, `pplr_get_data()` downloads the data from a  
169 server that hosts the database. This function can take as its first argument a `browse` object,  
170 a logical expression, either or both. The data downloaded from `popler` are in “long” form,  
171 meaning that each row of data reports a single measure of population size, and separate variables  
172 indicate the temporal and spatial replicate, taxa, etc. This format makes it easy to further subset  
173 downloaded datasets with the aim of visualization and analysis.

## 174 **Ancillary functions**

175 `popler` also provides three additional functions to open the url of the original dataset, un-  
176 pack covariates, and provide a citation for each dataset. First, the function `metadata_url()`  
177 launches the online study description in a web browser. Second, the `cov_unpack()` function  
178 transforms the `textttcovariates` variable into a data frame (which `pplr_get_data()` does not  
179 provide by default). Third, `pplr_citation()` generates a citation for the originators or each  
180 data set.

## 181 **Limitations and opportunities for development**

182 Working with raw, spatially replicated, and non-aggregated data provides key advantages in  
183 quantitative analyses of population dynamics which were a driving force behind the development  
184 of `popler`. However, users need to examine individual datasets and the associated online study  
185 descriptions to understand their peculiarities. Single datasets have unique idiosyncrasies that  
186 require vetting. For example, many datasets have gaps or changes in the sampling design during  
187 the length of the study, or the `covariates` variable can hold key information. Hence, we urge  
188 authors to consult the online documentation of the original datasets.

189 In the future, there are opportunities to increase the size of `popler` and expand its scope.  
190 First, because many of the studies included in `popler` are ongoing, there will be opportunities to

run regular updates aimed at including new observations in `popler`. Second, because our schema (Fig. 1) is very general, the database could be expanded to include population datasets outside of the LTER network. Third, it would be valuable to explicitly associate `popler`'s population-level data with environmental drivers, especially climate. It is our intention and hope that the resources provided by `popler` will advance ecological understanding of population dynamics within the LTER network, and more generally.

## Acknowledgements

We thank Trevor Drees and Michael Saucedo for assistance in database development. Support for database and package development was provided by the US National Science Foundation to TEXTM (DEB-1543651). This research was additionally supported by a Julian Huxley Faculty Fellowship from Rice University and a Faculty Research Grant awarded by the Committee on Research from the University of California, Santa Cruz (KZ). The LTER network is supported by the US National Science Foundation.

## Authors' contributions

AC, AB, KZ, MO, TEXTM designed and built the database. AC AB, KZ, BD, SM, and TEXTM designed and built the R package. AC and TEXTM led the writing of the manuscript. All authors contributed to manuscript drafts and gave final approval for publication.

## References

- W. A. Bechtold, P. L. Patterson, et al. *The enhanced forest inventory and analysis program: national sampling design and estimation procedures*, volume 80. US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina, 2005.
- J. T. Callahan. Long-term ecological research. *BioScience*, 34(6):363–367, 1984.
- S. A. Chamberlain and E. Szöcs. `taxize`: taxonomic search and retrieval in r. *F1000Research*, 2, 2013.

215 A. Compagnoni, A. J. Bibian, B. M. Ochocki, H. S. Rogers, E. L. Schultz, M. E. Sneek, B. D.  
216 Elderd, A. M. Iler, D. W. Inouye, H. Jacquemyn, and T. E. X. Miller. The effect of demographic  
217 correlations on the stochastic population dynamics of perennial plants. 86:480–494, 2016. ISSN  
218 0012-9615. doi: 10.1002/ecm.1228.

219 M. Dornelas, L. H. Antao, F. Moyes, A. E. Bates, A. Magorrra, D. Adam, et al. Biotime: a  
220 database of biodiversity time series for the anthropocene. *Global Ecology and Biogeography*,  
221 2018.

222 R. A. Fisher, C. D. Koven, W. R. Anderegg, B. O. Christoffersen, M. C. Dietze, C. E. Farrior,  
223 J. A. Holm, G. C. Hurtt, R. G. Knox, P. J. Lawrence, et al. Vegetation demographics in earth  
224 system models: A review of progress and priorities. *Global change biology*, 24(1):35–54, 2018.

225 S. E. Hampton, E. E. Holmes, L. P. Scheef, M. D. Scheuerell, S. L. Katz, D. E. Pendleton, and  
226 E. J. Ward. Quantifying effects of abiotic and biotic drivers on community dynamics with  
227 multivariate autoregressive (mar) models. *Ecology*, 94(12):2663–2669, 2013.

228 P. Inchausti and J. Halley. Investigating long-term ecological variability using the global popu-  
229 lation dynamics database. *Science*, 293(5530):655–657, 2001.

230 J. Knape and P. de Valpine. Are patterns of density dependence in the global population dy-  
231 namics database driven by uncertainty about population abundance? *Ecology letters*, 15(1):  
232 17–23, 2012.

233 A. K. Knapp and M. D. Smith. Variation among biomes in temporal dynamics of aboveground  
234 primary production. *Science*, 291(5503):481–484, 2001.

235 A. K. Knapp, M. D. Smith, S. E. Hobbie, S. L. Collins, T. J. Fahey, G. J. Hansen, D. A. Landis,  
236 K. J. La Pierre, J. M. Melillo, T. R. Seastedt, et al. Past, present, and future roles of long-term  
237 experiments in the lter network. *BioScience*, 62(4):377–389, 2012.

238 T. K. Kratz, L. A. Deegan, M. E. Harmon, and W. K. Lauenroth. Ecological variability in space  
239 and time: Insights gained from the us lter program. *AIBS Bulletin*, 53(1):57–67, 2003.

240 W. F. Laurance, F. Achard, S. Peedell, and S. Schmitt. Big data, big opportunities. *Frontiers*  
241 *in Ecology and the Environment*, 14(7):347–347, 2016.

242 J. M. Levine and J. HilleRisLambers. The importance of niches for the maintenance of species  
243 diversity. *Nature*, 461(7261):254, 2009.

244 D. B. Lindenmayer, G. E. Likens, A. Andersen, D. Bowman, C. M. Bull, E. Burns, C. R.  
245 Dickman, A. A. Hoffmann, D. A. Keith, M. J. Liddell, et al. Value of long-term ecological  
246 studies. *Austral Ecology*, 37(7):745–757, 2012.

247 J. Loh, R. E. Green, T. Ricketts, J. Lamoreux, M. Jenkins, V. Kapos, and J. Randers. The living  
248 planet index: using species population time series to track trends in biodiversity. *Philosophical*  
249 *Transactions of the Royal Society of London B: Biological Sciences*, 360(1454):289–295, 2005.

250 D. Medvigy, S. Wofsy, J. Munger, D. Hollinger, and P. Moorcroft. Mechanistic scaling of ecosys-  
251 tem function and dynamics in space and time: Ecosystem demography model version 2. *Journal*  
252 *of Geophysical Research: Biogeosciences*, 114(G1), 2009.

253 C. J. E. Metcalf and S. Pavard. Why evolutionary biologists should be demographers. *Trends in*  
254 *Ecology & Evolution*, 22(4):205–212, 2007.

255 D. Schimel, W. Hargrove, F. Hoffman, and J. MacMahon. Neon: A hierarchically designed  
256 national ecological network. *Frontiers in Ecology and the Environment*, 5(2):59–59, 2007.

257 K. R. Wilcox, A. T. Tredennick, S. E. Koerner, E. Grman, L. M. Hallett, M. L. Avolio, K. J.  
258 La Pierre, G. R. Houseman, F. Isbell, D. S. Johnson, J. M. Alatalo, A. H. Baldwin, E. W.  
259 Bork, E. H. Boughton, W. D. Bowman, A. J. Britton, J. F. Cahill, S. L. Collins, G. Du,  
260 A. Eskelinen, L. Gough, A. Jentsch, C. Kern, K. Klanderud, A. K. Knapp, J. Kreyling, Y. Luo,  
261 J. R. McLaren, P. Magonigal, V. Onipchenko, J. Prevéy, J. N. Price, C. H. Robinson, O. E.  
262 Sala, M. D. Smith, N. A. Soudzilovskaia, L. Souza, D. Tilman, S. R. White, Z. Xu, L. Yahdjian,  
263 Q. Yu, P. Zhang, and Y. Zhang. Asynchrony among local communities stabilises ecosystem  
264 function of metacommunities. *Ecology letters*, 20:1534–1545, Dec. 2017. ISSN 1461-0248. doi:  
265 10.1111/ele.12861.

Table 1: Variables used to store population or individual data in `popler`.

Variable	Description
<code>abundance_observation</code>	Measure of population abundance at a specific time and location. This variable measures abundance as a count, biomass, density, or cover. For individual data sets this variable is always equal to 1, because each attribute or set of attributes refer to a single individual.
<code>day</code>	Day of observation
<code>month</code>	Month of observation
<code>year</code>	Year of observation
<code>spatial_replicate_n</code>	The $n^{th}$ level of spatial replication, where <code>spatial_replicate_1</code> is the study site. <code>popler</code> accommodates up to five levels of spatial replication.
<code>treatment_type_n</code>	For datasets originating from an experimental study, the $n^{th}$ treatment. <code>popler</code> accommodates up to three treatments.
<code>covariates</code>	Ancillary observations that do not fall into the standard schema of <code>popler</code> .
<code>structure_type_n</code>	For individual data, these variables measure the $n^{th}$ attribute of individuals (identity, size, sex, status, stage). <code>popler</code> accommodates up to four structure types per dataset.

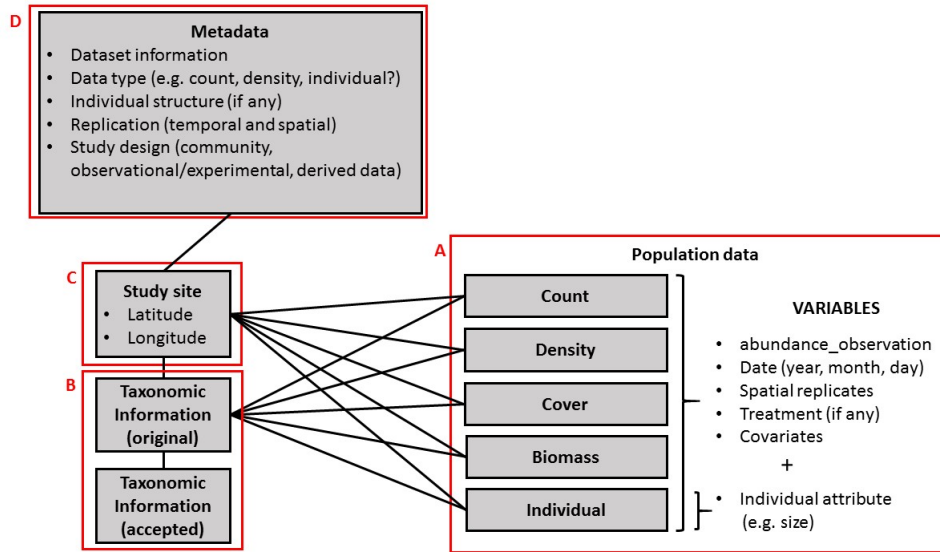


Figure 1: Schematic representation of the entity relationship diagram of the `popler` database. `popler` provides metadata on the studies that originated abundance data points (D). This metadata contains information on the unique identifiers of each study, on its design (observational or experimental), temporal and spatial replication. `popler` stores the latitude and longitude of the study site (C). Each abundance data point corresponds to a specific taxonomic unit (B). Finally, the time series of population data collected in a study can be of four different types (count, density, biomass, cover), or they may be individual data with attributes such as size or sex (A).

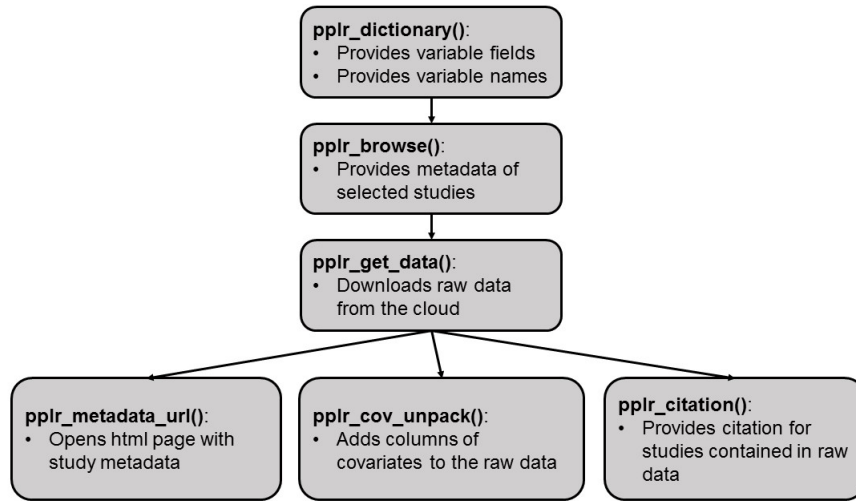


Figure 2: Suggested workflow when using the `popler` R package to interface with the homonymous online database. The function `pplr_dictionary()` refers to the variables of the metadata that describe the data sets contained in `popler`. `pplr_dictionary()` describes these variables and returns their possible values. This information advises which criteria to use when subsetting `popler`. The user can provide a criterion (that is, a logical statement) to browse the metadata, using `pplr_browse()`, or to download data using `pplr_get_data()`. Moreover, the output of `pplr_get_data()` (a data frame) can be the argument of three ancillary functions: `pplr_metadata_url()` opens the webpage containing the original dataset and their associated online metadata. `pplr_cov_unpack()` can be used to format the covariates contained in a raw data object into separate columns of a data frame. Finally, `pplr_citation()` provides a citation for the downloaded data set(s).



## 266 Appendix S1: Pre-processing `popler` data

267 Before uploading datasets into the online `popler` database, we combined datasets, transformed  
268 datasets from wide to long form, converted non-ASCII characters, and modified ambiguous study  
269 site names.

270 The variables of many datasets were contained in two or more separate files, which we com-  
271 bined in a single file. When the original dataset provided data in wide form, we transformed  
272 it into long form. In wide form datasets, abundance data associated with different species was  
273 stored in separate columns. `popler` stores these datasets in long form, whereby each row of  
274 abundance data is related to a specific taxonomic unit in the table containing taxonomic infor-  
275 mation (Fig. 1B). We converted all data in ASCII format, because the encoding of the database  
276 is the UTF-8. We often re-defined study site names to unambiguously associate them with one  
277 of the 26 LTER sites. Many site names are alphanumeric codes (e.g. “U1”) which can overlap  
278 across several LTER sites. Hence, we changed site names following a standard formula (namely,  
279 from “U1” to “site\_sbc\_U1”, where “sbc” refers to the Santa Barbara coastal LTER site).

280 In a handful of cases, we removed single data rows from the original dataset. These data  
281 rows were associated with two types of typos in the original dataset. First, some abundance  
282 observations were not associated with a time of observation. We removed this data because  
283 `popler` can only accommodate population information associated with a time of observation.  
284 Second, a handful of abundance data points were clear typos (e.g. the letter “l” instead of a  
285 numeric value). We substituted these data points with a missing value (NULL in the database).  
286 We uploaded these pre-processed datasets in the `popler` database through a Graphic User  
287 Interface developed in Python using libraries `panda` and `pyqt5`.

Table S1: Taxonomic variables contained in the popler table on original taxonomic information.

Variable
sppcode
kingdom
subkingdom
infrakingdom
superdivision
division
subdivision
superphylum
phylum
subphylum
class
subclass
order
family
genus
species
common_name

Table S2: Metadata variables used to describe the datasets stored in popler.

Variable	Description
proj_metadat_key	Unique ID
lter_project_key	ID of LTER site
lter_project_key	ID of LTER site
title	Title of study
samplingunits	Unit of measure (if any) referred to population data.
datatype	Data type: count, biomass, cover, density, and individual. These correspond to the tables in Fig. 1A.
structured_type_n	If individual data, this shows what type of structure is stored. A study can contain up to $n = 4$ types of structure.
structured_type_n_units	Unit of measure (if any) referred to structure data.
studystartyr	Start year of the study
studyendyr	End year of the study
duration_years	Duration of the study in years
samplefreq	Frequency of population census
studytype	Whether study is observational or experimental
community	Whether study includes single taxon (community = F) or multiple taxa (community = T)
spatial_replication_level_n_extent	Extent of spatial replication level number $n$ . A dataset can have up to 5 replication levels.
spatial_replication_level_n_extent_units	Unit of spatial extent of the $n$ spatial replication level.
spatial_replication_level_n_label	Label of the spatial replication level (e.g. transect, plot, quadrat, ect.). The label of spatial replication level 1 is "site".
spatial_replication_level_n_number_of_unique_reps	The number of unique replicates for the $n$ th level of spatial replication.
treatment_type_n	The type of treatment (e.g. resource manipulation). A study can contain up to $n = 3$ treatments.
control_group	If study is experimental, this shows the field(s) that identify the control replicate.
derived	Is population size data raw, or is it derived (e.g. it is aggregated)?
authors	Author(s) of the original dataset
authors_contact	Email address(es) of the author(s) associated with the original dataset.
metalink	url of the original dataset
knbid	Knowledge Network for Biocomplexity identifier.

Table S3: LTER identification acronyms and their meaning as used in the popler database.

Variable	LTER name
AND	Andrew Forest LTER
ARC	Arctic LTER
BES	Baltimore Ecosystem Study
BNZ	Bonanza Creek LTER
CAP	Central Arizona - Phoenix LTER
CCE	California Current Ecosystem LTER
CDR	Cedar Creek Ecosystem Science Reserve LTER
CWT	Coweeta LTER
FCE	Florida Coastal Everglades LTER
GCE	Georgia Coastal Ecosystems LTER
HBR	Hubbard Brook LTER
HFR	Harvard Forest LTER
JRN	Jornada Basin LTER
KBS	Kellogg Biological Station LTER
KNZ	Konza Prairie LTER
LNO	LTER Network Office
LUQ	Luquillo LTER
MCM	McMurdo Dry Valleys LTER
MCR	Moorea Coral Reef LTER
NCO	LTER Network Communications Office
NTL	North Temperate Lakes LTER
NWT	Niwot Ridge LTER
PAL	Palmer Antarctica LTER
PIE	Plum Island Ecosystems LTER
SBC	Santa Barbara Coastal LTER
SEV	Sevilleta LTER
SGS	Shortgrass Steppe LTER
VCR	Virginia Coastal Reserve LTER