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

Aldo Compagnoni*[a,b,c], Andrew J. Bibian[a], Brad M. Ochocki[a], Sam Levin[b,c], Kai Zhu[d] and Tom E.X. Miller[a]

[a]Department of BioSciences, Program in Ecology and Evolutionary Biology, Rice University, 6100 Main St, MS-170, Houston, TX 77005

[b]Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany

[c]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

[d]Department of Environmental Studies, University of California, Santa Cruz, CA 95064, USA

Running headline: The `popler` database and R package

*aldo.compagnoni@gmail.com

# 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 Stated 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 305 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("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

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

# 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

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

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

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

## The `popler` database

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

observations.

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

The popler database currently contains data from 305 studies (122 of which are experimental) representing 4377 cumulative years of observations with a mean study duration of 14.35 years. popler contains data from 665 plant species, 382 animal species, and 1 fungal species.

## Population data

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

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

popler contains both observational and experimental studies. Experimental datasets contain information on one or more experimental treatments. Popler accommodates information on up to three experimental treatments, identified by three numbered treatment_type variables

(Table 1).

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

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

## Taxonomic information

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

(Chamberlain and Szöcs, 2013).

### Study site

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

### Metadata

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

## The `popler` package

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

## The `pplr_dictionary()` function

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

## The `pplr_browse()` function

Once the user is familiar with the meaning and content of the variables that define `popler` datasets, they are ready to dig deeper using `pplr_browse()` (Fig. 2). Running `pplr_browse()` without arguments provides the metadata from the entire contents of the database. This will be a $305 by 20$ data frame, with each row corresponding to a study and each column corresponding to a variable defined by `pplr_dictionary()`.

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

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

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

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

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

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

### The `pplr_get_data()` function

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

### Ancillary functions

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

## Limitations and opportunities for development

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

In the future, there are opportunities to increase the size of `popler` and expand its scope. 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. Thus, 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, Maurizio Compagnoni for assistance in database management, and Scott Chamberlain for developing the API to query the online database. Support for database and package development was provided by the US National Science Foundation to TEXM (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, TEXM designed and built the database. AC AB, KZ, BD, SM, and TEXM designed and built the R package. AC and TEXM led the writing of the manuscript. All authors contributed to manuscript drafts and gave final approval for publication.

## Data Availability

The `popler` R package is publicly available at https://github.com/ropensci/popler.

## 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

K. R. Wilcox, A. T. Tredennick, S. E. Koerner, E. Grman, L. M. Hallett, M. L. Avolio, K. J. La Pierre, G. R. Houseman, F. Isbell, D. S. Johnson, J. M. Alatalo, A. H. Baldwin, E. W. Bork, E. H. Boughton, W. D. Bowman, A. J. Britton, J. F. Cahill, S. L. Collins, G. Du, A. Eskelinen, L. Gough, A. Jentsch, C. Kern, K. Klanderud, A. K. Knapp, J. Kreyling, Y. Luo, J. R. McLaren, P. Megonigal, V. Onipchenko, J. Prevéy, J. N. Price, C. H. Robinson, O. E. Sala, M. D. Smith, N. A. Soudzilovskaia, L. Souza, D. Tilman, S. R. White, Z. Xu, L. Yahdjian, Q. Yu, P. Zhang, and Y. Zhang. Asynchrony among local communities stabilises ecosystem

267 function of metacommunities. *Ecology letters*, 20:1534–1545, Dec. 2017. ISSN 1461-0248. doi:
268 10.1111/ele.12861.

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

| Variable | Description |
|---|---|
| `abundance_observation` | 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. |
| `day` | Day of observation |
| `month` | Month of observation |
| `year` | Year of observation |
| `spatial_replicate_n` | The $n^{th}$ level of spatial replication, where `spatial_replicate_1` is the study site. `popler` accommodates up to five levels of spatial replication. |
| `treatment_type_n` | For datasets originating from an experimental study, the $n^{th}$ treatment. `popler` popler accommodates up to three treatments. |
| `covariates` | Ancillary observations that do not fall into the standard schema of `popler`. |
| `structure_type_n` | For individual data, these variables measure the $n^{th}$ attribute of individuals (identity, size, sex, status, stage). `popler` accommodates up to four structure types per dataset. |

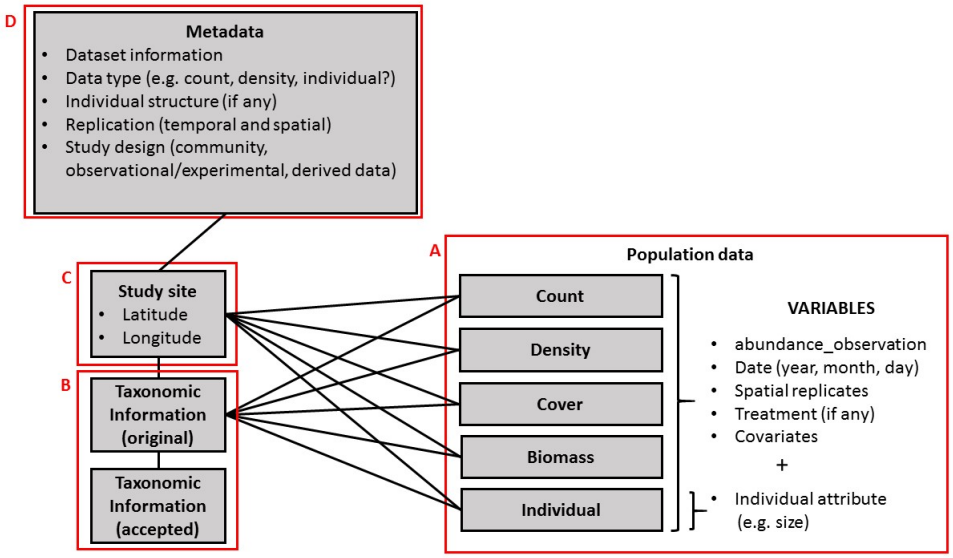


Figure 1: Schematic representation of the entity relationship diagram of the `popler` database. `popler` provides metadata on the studies that originated abundance data points (D). This metadata contains information on the unique identifiers of each study, on its design (observational or experimental), 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).

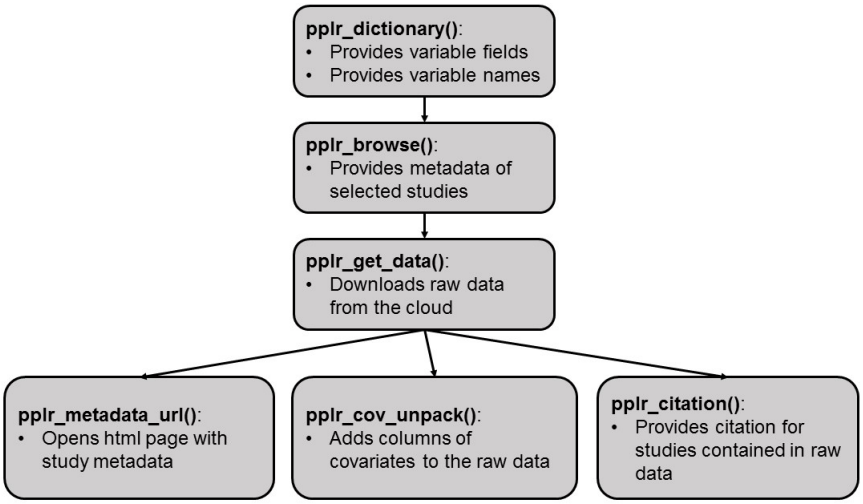


Figure 2: Suggested workflow when using the `popler` R package to interface with the homonymous online database. The function `pplr_dictionary()` refers to the variables of the metadata that describe the data sets contained in `popler`. `pplr_dictionary()` describes these variables and returns their 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_gettunderscore 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).