8-3-2018

# MSB-ECA: Phylogenetically-informed modeling of the regional context of community assembly

William D. Pearse
*Utah State University*, will.pearse@usu.edu



## Recommended Citation

Pearse, W. D. (2018). MSB-ECA: Phylogenetically-informed modeling of the regional context of community assembly. Utah State University. https://doi.org/10.15142/T3RS84

# Data management plan

Data release and management form a core part of objective 1 of this proposal. The last few years have seen transformative developments in data management tools, and this proposal will leverage these to ensure that all my data is openly released and easy to access for all. Becoming part of established and well-used infrastructure is the best way to ensure that data is accessible now and in the future: developing or experimenting with novel data housing solutions is not the goal of this project. Below, I describe in detail data management plans for all functional trait and phylogenetic data, alongside plans for the storage and release of computational tools. Please note that NEON are releasing all data associated with the species at each site, and so that is not discussed in this document. There are no publishing fees associated with any of the platforms I describe below.

## Images of specimens

A key component of objective 1 is the digitization of specimens using dedicated equipment at Utah State University (see letters of collaboration). These images will be uploaded, along with the appropriate meta-data, to Integrated Digitized Biocollections (iDigBio; https://www.idigbio.org/), which is the repository supported by the National Resource for Advancing Digitization of Biodiversity Collections (funded by the National Science Foundation). iDigBio will ensure that the data are then safely housed and distributed freely.

Details of:
**Equipment for capture.** All plant herbarium specimens will be digitized using color flatbed scanners with a minimum resolution of 300 ppi. All beetle and tick specimens will be digitized using a Keyence VHX-5000 digital microscope at the USDA-ARS Pollinating Insects Research Unit on the campus of Utah State University.
**Files for release.** Approximately ~8600 scanned 300 ppi 24-bit TIFF-format flatbed-scanned plant images, ~1450 photographed DNG-format beetle images, and ~500 (estimated) DNG-format tick images will be released. These file formats adhere to the recording and archival standards set by iDigBio.
**Meta-data format.** Meta-data including the location, species identity, provenance, and capture method of all specimens will be recorded using the Darwin Core meta-data standard. This file format adheres to the standards set by iDigBio.
**Release timeline.** Images and their meta-data will be uploaded to iDigBio within one year of capture. They will, therefore, likely be publicly accessible before the functional trait data derived from them are available (see below).
**Archival and long-term availability.** iDigBio is an NSF-funded initiated intended to support the storage and distribution of specimen images. It current holds over 15 million images, and is in my opinion the most stable infrastructure for the foreseeable future.
**Short-term backup.** Before public release, data will be stored using three different methods: on the hard-drive of the post-doc's computer, on an external hard-drive stored away from the post-doc's computer, and on the cloud-based Dropbox servers.
**Release conditions.** These data will be released under the CC BY (Attribution) license, which

allows users to copy, transmit, reuse, remix, and adapt content, and requires attribution of the source of the content (see http://creativecommons.org/licenses/by/4.0/)

## Functional trait databases

All trait data will be released through the FigShare platform (https://figshare.com/). Used by a number of publishers (including Wiley, Springer Nature, Taylor and Francis, and PLoS), I have experience using this platform (*e.g.*, http://dx.doi.org/10.6084/m9.figshare. 979288) and am confident data will be available through it for years to come. Data without provenance is of limited use: for this reason, I will also be archiving the data pipelines used to generate each database (see also "*Computation tools*" below).

Details of:
**Files for release.** Tab-delimited text files with continuous and categorical traits for the plants, beetles and ticks studied in this proposal.
**Meta-data format.** Data will be described using Ecological Metadata Language (EML). This format is based on, and readable as, XML, and was developed in association with the Ecological Society of America (https://knb.ecoinformatics.org/#external//emlparser/docs/index.html).
**Release timeline.** Data will be released as research publications making use of the data are submitted for review or at the end of the funded project, whichever is the soonest.
**Archival and long-term availability.** FigShare is used by many publishers (see above), and as such I am confident it will remain a freely and widely accessible source of data for many years to come
**Short-term backup.** Before public release, data will be stored using three different methods: on the hard-drive of the post-doc's computer, on an external hard-drive stored away from the post-doc's computer, and on the cloud-based Dropbox servers.
**Release conditions.** These data will be released under the CC0 license, which is recommended by FigShare for datasets since they contain facts that are, in of themselves, not protected by law. This removes the legal, but not the moral, obligation to cite data when it is used, and gives no restrictions on the further use of data by other scientists.

## DNA sequences and phylogenies

All DNA sequences are being collected and distributed by NEON, and so are outside the scope of this proposal. The phylogenies produced as part of objective 1 will be released in the online repository TreeBASE (https://treebase.org/). TreeBASE has established recommendations for the storage and sharing of data produced as part of any National Science Foundation grant (see https://treebase.org/treebase-web/dataMan.html), and I will comply with all the meta-data requirements they outline (including depositing DNA alignments used to produce the phylogenies).

Details of:
**Files for release.** Plant and beetle phylogenies. Note: I will make use of, but not be making the first release of, DNA data collected by NEON.

**Meta-data format.** Following TreeBASE's suggestions, NeXML data that will contain the phylogenies themselves, the DNA alignments used to create them, NCBI taxon identification numbers, and the GenBank accession numbers for the sequences used.
**Release timeline.** Data will be released as research publications making use of the data are submitted for review, or at the end of the funded project, whichever is the soonest.
**Archival and long-term availability.** TreeBASE has been in existence for over twenty years, and so is likely to continue to operate for the foreseeable future. It is the *de facto* standard for freely sharing phylogenetic information.
**Short-term backup.** Before public release, data will be stored using three different methods: on the hard-drive of the post-doc's computer, on an external hard-drive stored away from the post-doc's computer, and on the cloud-based Dropbox servers.
**Release conditions.** All TreeBASE data are freely available.

## Analytical software

I currently maintain phylogenetics software on my GitHub website (github.com/willpearse/phyloGenerator2) and Phylogenetic Generalized Linear Mixed Modeling software on the central *R* server (CRAN; cran.r-project.org/package=pez). Updates developed for these pieces of software will be released through these existing open access routes.

Details of:
**Files for release.** All source code necessary to run, and where necessary compile, the software. Unit tests necessary to check the integrity of the code.
**Meta-data format.** Language-specific help files, vignettes, and online documentation, and videos demonstrating use (*e.g.*, willpearse.github.io/phyloGenerator/guide.html#screencasts).
**Release timeline.** Constantly updated throughout the project.
**Archival and long-term availability.** CRAN (for R code) has been freely distributing code for 19 years, and I therefore consider it a safe release platform. GitHub has become the *de facto* industry standard for code release, and after five years successfully distributing code with it I am confident it will continue to do so for the foreseeable future.
**Short-term backup.** Before public release, code will be stored using three different methods: on the hard-drive of the post-doc's computer, and on the cloud-based GitHub servers.
**Release conditions.** All code will be released under the GNU GPL v3 license, which permits free copy, distribution, and use.

## Data processing pipelines

I have experience maintaining database cleaning and management pipelines on my GitHub website (github.com/willpearse/urbanHomogenizationDB). I will continue to use GitHub to store this data, and will additionally store all pipelines related to the cleaning and production of data within the FigShare repository associated with the data itself (see above).

Details of:
**Files for release.** All source code necessary to generate the databases (see "*Functional trait*

*databases*") when given raw input data (see also "*Images of specimens*"). Unit tests necessary to check the integrity of the code.

**Meta-data format.** The outputted data will be described using Ecological Metadata Language (EML). This format is based on, and readable as, XML, and was developed in association with the Ecological Society of America (https://knb.ecoinformatics.org/#external//emlparser/docs/index.html). The code will be described in language-specific help files, vignettes, and online documentation and videos demonstrating use (*e.g.*, http://willpearse.github.io/phyloGenerator/guide.html#screencasts).

**Release timeline.** Code will be constantly updated and released during the project on GitHub, but the final version associated with the databases (see "*Functional trait databases*") will be archived on FigShare along with the databases themselves.

**Archival and long-term availability.** FigShare is used by many publishers (see above), and as such I am confident it will remain a freely and widely accessible source of data for many years to come. I will be using GitHub for the intermediary versions of the pipelines (not the data themselves or the final versions); this will allow others to make use of the software underlying the database much faster.

**Short-term backup.** Before public release, code will be stored using three different methods: on the hard-drive of the post-doc's computer, and on the cloud-based GitHub servers.

**Release conditions.** All code will be released under the GNU GPL v3 license, which permits free copy, distribution, and use.