ecocomDP: A data design pattern and R package to facilitate FAIR biodiversity data for ecological synthesis

Eric R Sokol ‡, §

‡ Battelle, National Ecological Observatory Network (NEON), Boulder, CO, United States of America
§ Institute of Arctic and Alpine Research (INSTAAR), University of Colorado Boulder, Boulder, CO, United States of America

Corresponding author: Eric R Sokol (esokol@battelleecology.org)

Abstract

Two programs that provide high-quality long-term ecological data, the Environmental Data Initiative (EDI) and the National Ecological Observatory Network (NEON), have recently teamed up with data users interested in synthesizing biodiversity data, such as ecological synthesis working groups supported by the US Long Term Ecological Research (LTER) Network Office, to make their data more Findable, Interoperable, Accessible, and Reusable (FAIR). To this end:

1. we have developed a flexible intermediate data design pattern for ecological community data (L1 formatted data in Fig. 1, see Fig. 2 for design details) called "ecocomDP" (O'Brien et al. 2021), and
2. we provide tools to work with data packages in which this design pattern has been implemented.

The ecocomDP format provides a data pattern commonly used for reporting community level data, such as repeated observations of species-level measures of biomass, abundance, percent cover, or density across multiple locations. The ecocomDP library for R includes tools to search for data packages, download or import data packages into an R (programming language) session in a standard format, and visualization tools for data exploration steps that are recommended for data users prior to any cross-study synthesis work. To date, EDI has created 70 ecocomDP data packages derived from their holdings, which include data from the US Long Term Ecological Research (US LTER) program, Long Term Research in Environmental Biology (LTREB) program, and other projects, which are now discoverable and accessible using the ecocomDP library. Similarly, NEON data products for 12 taxonomic groups are discoverable using the ecocomDP search tool. Input from data users provided guidance for the ecocomDP developers in mapping the NEON data products to the ecocomDP format to facilitate interoperability with the ecocomDP data packages available from the EDI repository. The standardized data design pattern allows common data visualizations across data packages, and has the potential to facilitate the

development of new tools and workflows for biodiversity synthesis. The broader impacts of this collaboration are intended to lower the barriers for researchers in ecology and the environmental sciences to access and work with long-term biodiversity data and provide a hub around which data providers and data users can develop best practices that will build a diverse and inclusive community of practice.

Keywords

long-term data, US LTER, NEON, data discovery, data interoperability, community ecology

Presenting author

Eric R. Sokol

Presented at

TDWG 2021

Acknowledgements

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

Conflicts of interest

References

- O'Brien M, Smith C, Sokol E, Gries C, Lany N, Record S, Castorani MN (2021) ecocomDP: A flexible data design pattern for ecological community survey data. Ecological Informatics 64 https://doi.org/10.1016/j.ecoinf.2021.101374

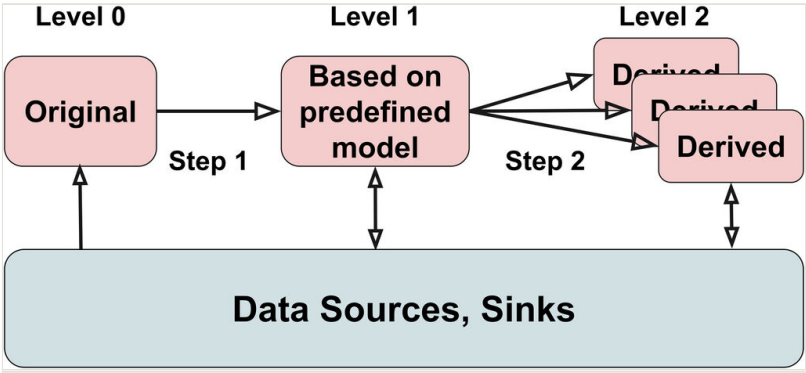

Figure 1.

Generalized flow of data in ecological synthsis. Level 0 (L0) are incoming, original data, ideally, already archived in the repository with complete metadata and contributed by those close to the research. Level 1 (L1) data packages (also in the repository) are formatted according to a predefined model, in this case, ecocomDP. Researchers are able to use L1 as inputs with its code to speed their analyses and generate Level 2 (L2) data. An archive of the L2 data package in the same repository is recommended. Data sources and sinks may be a repository (e.g., EDI) another data provider (e.g., NEON) or aggregator (e.g., GBIF). Reproduced from O'Brien et al. (2021).

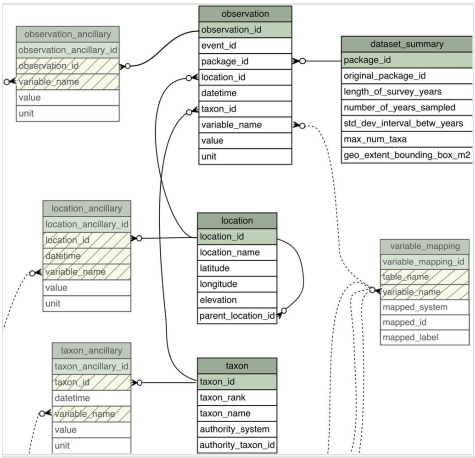

Figure 2.

The ecocomDP model shown with relational database notation for foreign keys and relationships (e.g, lines ending in crows-foot indicate 1:many relationships). Semi-transparent tables are optional. Medium green fields in each table are the primary key. Yellow/hashed fields are a combined unique constraint. IDs (suffixed, "_id"), must be unique within a table, as in an relational database. Full documentation can be found here. Reproduced from O'Brien et al. (2021).