Leveraging ecocomDP as a Flexible Intermediate Data Pattern to Expose NEON Biodiversity Data in GBIF

Eric R Sokol[‡], Colin A Smith[§], Margaret O'Brien[|]

‡ National Ecological Observatory Network (NEON), Battelle, Boulder, CO, United States of America
§ Center for Limnology, University of Wisconsin, Madison, WI, United States of America
| Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA, United States of America

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

Abstract

The Environmental Data Initiative (EDI) and the National Ecological Observatory Network (NEON) have been developing a flexible intermediate data design pattern for ecological community data called "ecocomDP", which is intended to promote FAIR data principles. Specifically, this effort will enhance the discoverability of and access to biodiversity data from NEON and EDI data holdings, including data from the United States Long Term Ecological Research (USLTER) program (O'Brien et al. 2021). The ecocomDP data model is applied in the ecocomDP R (programming language) library, which provides tools for independent researchers to format their data following the ecocomDP standard, as well as tools to search and visualize data from NEON and EDI data holdings in their R environment. The flexibility of the ecocomDP data model allows for much of the ancillary data associated with observation events to be preserved. Here we describe a modular workflow that is under development to expose ecocomDP-formatted data packages in the Global Biodiversity Information Facility (GBIF) data portal (Fig. 1). Specifically, we highlight an effort to apply this workflow to create a pipeline to convert and submit NEON biodiversity data products to GBIF.

EDI now has more than 70 data packages reformatted to the ecocomDP model, and has nearly finished developing a conversion of that intermediate format to a Darwin Core Archive (DwC-A, event core) format (Wieczorek et al. 2012) for submission to GBIF. This workflow takes advantage of EDI's dataset subscription service, which triggers creation of an updated DwC-A when an original dataset is revised. Because ecocomDP provides a standardized input to this submission process, any data package in the ecocomDP format can be exposed in GBIF through this workflow. Thus, we are working to leverage the EDI-managed conversion and submission process to expose NEON data in GBIF, which is possible because of the existing mappings of NEON data products to ecocomDP (Li et al. 2022). This will include data products representing terrestrial and aquatic organisms (Table 1) from all NEON sites, spanning the entire United States.

The overall goal of this effort is to provide an automated, modular workflow with complete provenance to submit NEON and EDI datasets to GBIF, built in such a way that datasets can be properly updated as new samples are collected and the data are published. The development of such a submission pipeline will provide a standardized process to expose biodiversity data from two continental scale networks, NEON and the U.S. National Science Foundation's Long-term Ecological Research network in GBIF. Further, the modularity of the workflow will allow independent researchers to adapt tools developed in this effort for their data archiving and publishing needs.

Keywords

National Ecological Observatory Network, US Long Term Ecological Research program, USLTER, FAIR data

Presenting author

Eric R. Sokol

Presented at

TDWG 2022

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

- Li D, Record S, Sokol E, Bitters M, Chen M, Chung YA, Helmus M, Jaimes R, Jansen L, Jarzyna M, Just M, LaMontagne J, Melbourne B, Moss W, Norman KA, Parker S, Robinson N, Seyednasrollah B, Smith C, Spaulding S, Surasinghe T, Thomsen S, Zarnetske P (2022) Standardized NEON organismal data for biodiversity research. Ecosphere 13 (7). https://doi.org/10.1002/ecs2.4141
- 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 (101374). https://doi.org/10.1016/j.ecoinf.2021.101374

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715

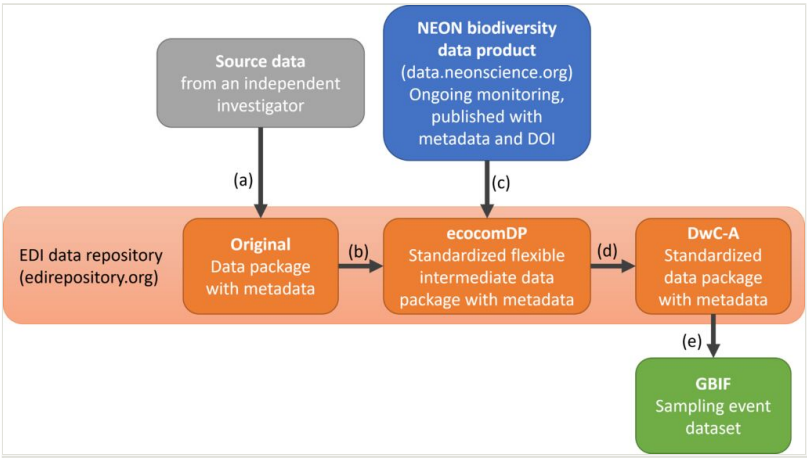

Figure 1.

Workflow to format and submit data to GBIF using ecocomDP as a flexible intermediate data format. All data packages stored in the EDI data repository are assigned digital object identifiers (DOIs), version controlled, and accessible through the EDI data portal. Investigators can (a) submit biodiversity datasets to the EDI repository for publication, which can then be (b) converted to the ecocomDP format using functions in the ecocomDP R library. NEON biodiversity data products can be (c) converted to the ecocomDP format using mappings available in the ecocomDP R library, and we are developing a process to submit the converted data to the EDI data repository. The ecocomDP data packages can then be (d) converted to Darwin Core Archives (DwC-A) that are stored in the EDI repository, which are then (e) submitted to GBIF as sampling event datasets.

Table 1.

NEON biodiversity datasets available in the ecocomDP format.

Taxonomic group	NEON data product ID	DOI for 2022 data release
Breeding land birds	DP1.10003.001	https://doi.org/10.48443/88sy-ah40
Ground beetles	DP1.10022.001	https://doi.org/10.48443/xgea-hw23
Herptile bycatch from ground beetle sampling	DP1.10022.001	https://doi.org/10.48443/xgea-hw23
Small mammals	DP1.10072.001	https://doi.org/10.48443/h3dk-3a71
Mosquitoes	DP1.10043.001	https://doi.org/10.48443/c7h7-q918
Terrestrial plants	DP1.10058.001	https://doi.org/10.48443/pr5e-1q60
Ticks	DP1.10093.001	https://doi.org/10.48443/7jh5-8s51
Tick pathogens	DP1.10092.001	https://doi.org/10.48443/nygx-dm71
Fishes	DP1.20107.001	https://doi.org/10.48443/7p84-6j62
Macroinvertebrates	DP1.20120.001	https://doi.org/10.48443/gn8x-k322
Microalgae	DP1.20166.001	https://doi.org/10.48443/g2k4-d258
Zooplankton	DP1.20219.001	https://doi.org/10.48443/150d-yf27