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

Eric R. Sokol[1*], Colin A. Smith[2], and Margaret O'Brien[3]

(1) National Ecological Observatory Network (NEON), Battelle, Boulder, CO, USA, *Email: esokol@battelleecology.org; (2) Center for Limnology, University of Wisconsin, Madison, WI, USA; (3) Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA, USA

Environmental Data Initiative
- Create . Package . Archive . Discover . Reuse -

NATIONAL SCIENCE FOUNDATION LTER NETWORK LONG TERM ECOLOGICAL RESEARCH

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 (LTER) program (O'Brien et al. 2021). The ecocomDP data model is applied in the ecocomDP R (programming language) package, which provides tools to:

- Format data following the ecocomDP standard
- Search the ecocomDP data catalog
- Import ecocomDP-formatted datasets into an R working environment
- Visualize and explore ecocomDP datasets (e.g., spatial and temporal replication, taxonomic resolution, etc.)

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 (**Figure 1**). Specifically, we highlight an effort to apply this workflow to create a pipeline to convert and submit NEON biodiversity data products to GBIF.

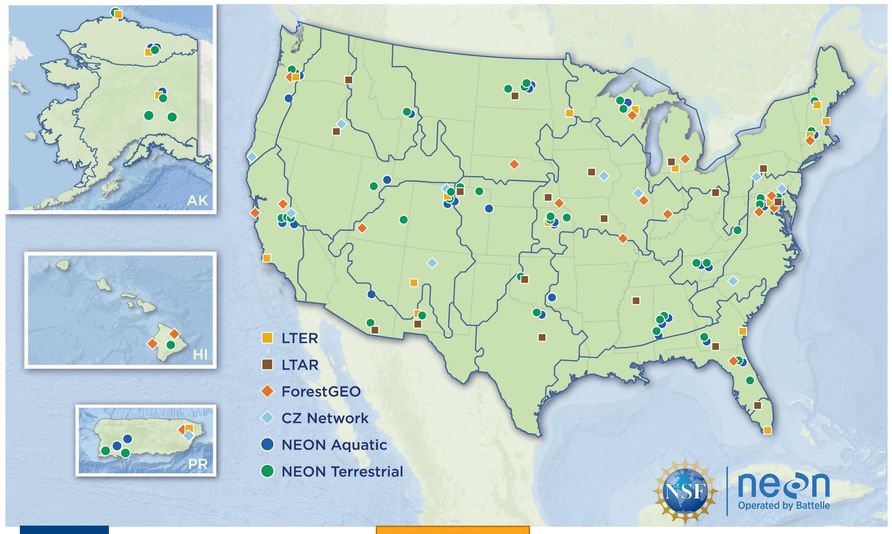
Source data

EDI now has more than 70 data packages, largely representing biodiversity data from LTER sites, reformatted to the ecocomDP model. Twelve NEON data products have been mapped to ecocomDP (Li et al. 2022). These include data products representing terrestrial and aquatic organisms (**Table 1**) from all NEON sites, spanning the entire United States (**Figure 1A**).

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

(A) Map of NEON and other co-located research sites



(B) Proposed data pathways to GBIF

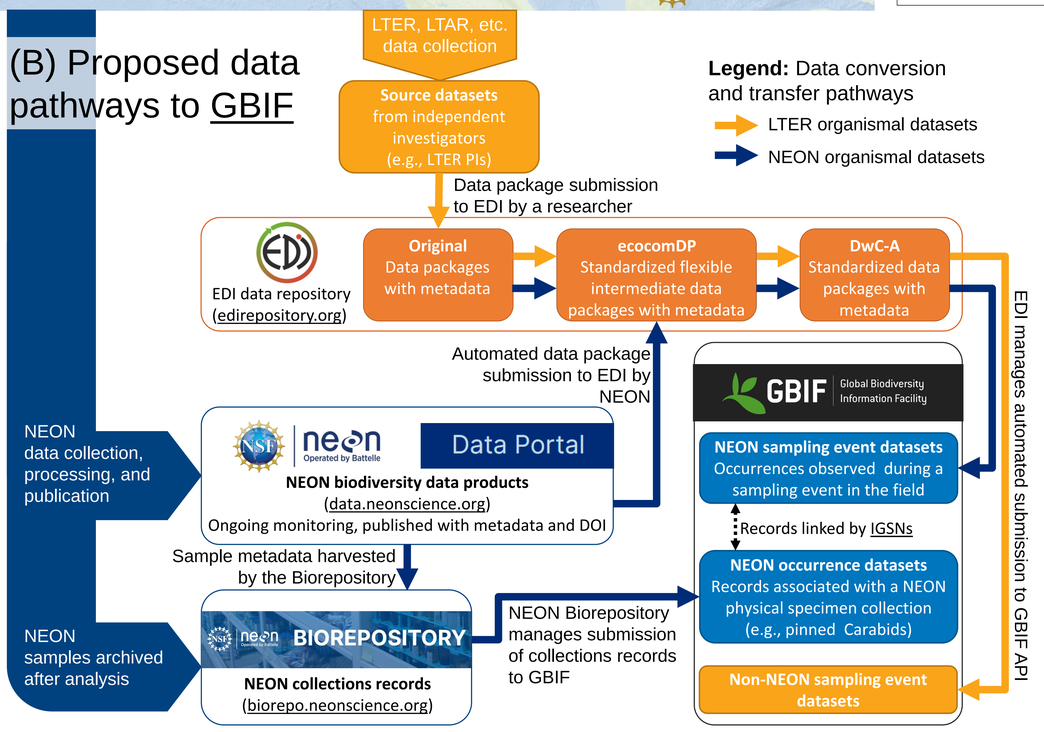


Figure 1. Overview of the proposed workflow to submit biodiversity data from NEON and EDI holdings to GBIF; **(A)** map of NEON, US LTER, and other long-term monitoring sites, where potential source data are collected; **(B)** proposed data submission pathways.

Proposed GBIF submission workflows

EDI has nearly finished developing a workflow to convert ecocomDP data packages to the Darwin Core Archive (DwC-A, event core) format (Wieczorek et al. 2012), which can then be submitted to GBIF. This workflow can be applied to both **LTER** and **NEON** organismal datasets (**Fig 1B)**.

Pathway for LTER and single-PI dataset submission to GBIF:

1. Submit data package to EDI repository
2. Convert to ecocomDP formatted derived data package
3. Convert to DwC-A formatted derived data package
4. Submit to GBIF

Pathway for NEON organismal dataset submission to GBIF:

1. Use NEON pre-built mappings in the ecocomDP R package to convert NEON organismal data products to ecocomDP format and submit to EDI
2. Convert to DwC-A formatted derived data package
3. Submit to GBIF. **NOTE:** The NEON Biorepository already exposes NEON collections data for archived samples in GBIF. Collections records will be linked to field observation records using IGSNs where appropriate.

Future directions

Working toward automation

These workflows can be automated because EDI's dataset subscription service triggers creation of an updated DwC-A when an upstream original dataset is revised. As this workflow is developed, EDI will work with data providers to determine which steps in the process can and should be automated for different use cases.

Generality

Because ecocomDP provides a standardized input to the GBIF submission process, any data package in the ecocomDP format in the EDI data repository can be exposed in GBIF through this workflow. Further, the modularity of the workflow will allow independent researchers to adapt tools developed in this effort for their data archiving and publishing needs.

Cross-network data harmonization

This submission pipeline will provide a standardized process to expose biodiversity data from two continental scale networks, NEON and the US LTER, in GBIF.

https://doi.org/10.3897/biss.6.93915

Creative Commons Attribution License (CC BY 4.0)