2021

# ecocomDP: A flexible data design pattern for ecological community survey data

Margaret O'Brien

Colin A. Smith

Eric R. Sokol

Corinna Gries

Nina Lany

*See next page for additional authors*

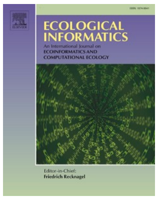
## Authors

Margaret O'Brien, Colin A. Smith, Eric R. Sokol, Corinna Gries, Nina Lany, Sydne Record, and Max C. N. Castorani

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



# ecocomDP: A flexible data design pattern for ecological community survey data

Margaret O'Brien [a,*], Colin A. Smith [b], Eric R. Sokol [c,d], Corinna Gries [b], Nina Lany [e], Sydne Record [f], Max C.N. Castorani [g]

[a] *Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA 93106, United States of America*
[b] *Center for Limnology, University of Wisconsin, Madison, WI 53706, United States of America*
[c] *Battelle, National Ecological Observatory Network (NEON), Boulder, CO 80301, United States of America*
[d] *Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, CO 80309, United States of America*
[e] *U.S. Forest Service, 271 Mast Road, Durham, NH 03824, United States of America*
[f] *Department of Biology, Bryn Mawr College, Bryn Mawr, PA 19010, United States of America*
[g] *Dept. of Environmental Sciences, University of Virginia, Charlottesville, VA 22904, United States of America*

ARTICLE INFO

*Keywords:*
Ecological community survey
Data harmonization
Workflow
ecocomDP
NEON
LTER
LTREB

ABSTRACT

The idea of harmonizing data is not new. Decades of amassing data in databases according to community standards - both locally and globally - have been more successful for some research domains than others. It is particularly difficult to harmonize data across studies where sampling protocols vary greatly and complex environmental conditions need to be understood to apply analytical methods correctly. However, a body of long-term ecological community observations is increasingly becoming publicly available and has been used in important studies. Here, we discuss an approach to preparing harmonized community survey data by an environmental data repository, in collaboration with a national observatory. The workflow framework and repository infrastructure are used to create a decentralized, asynchronous model to reformat data without altering original data through cleaning or aggregation, while retaining metadata about sampling methods and provenance, and enabling programmatic data access. This approach does not create another data 'silo' but will allow the repository to contribute subsets of available data to a variety of different analysis-ready data preparation efforts. With certain limitations (e.g., changes to the sampling protocol over time), data updates and downstream processing may be completely automated. In addition to supporting reuse of community observation data by synthesis science, a goal for this harmonization and workflow effort is to contribute these datasets to the Global Biodiversity Information Facility (GBIF) to increase the data's discovery and use.

## 1. Introduction

Primary environmental research data are being made publicly available based on two main premises. First, the practice will make research more transparent and back up results, and second, it will enable reusing the data in more than one research project (Heffernan et al., 2014). Specifically, the combination of many local-scale research results may reveal broader patterns, drivers, trajectories, and predictions of ecological systems, particularly in response to the current rapid and unprecedented environmental changes (Levy et al., 2014). Many research communities have recognized this potential and data repositories like the Environmental Data Initiative (EDI, https://EnvironmentalDataInitiative.org) hold thousands of diverse primary datasets from research studies in the ecological sciences. However, these data, although publicly available, still remain mostly locked away by their varied sampling methodologies, idiosyncratic formatting and non-standardized terminology. Furthermore, these data can only be reused when the environmental context in which they were collected is fully understood and accounted for in the analytical approaches (Welti et al., 2021).

Given this situation, primary research datasets in ecology are often not easily combined or synthesized. Comprehending sampling and environmental conditions, resolving terminology, formatting, and aggregating data generally takes a large portion of research time (Lohr,

[*] Corresponding author.
*E-mail addresses:* margaret.obrien@ucsb.edu (M. O'Brien), castorani@virginia.edu (M.C.N. Castorani).

https://doi.org/10.1016/j.ecoinf.2021.101374
Received 28 December 2020; Received in revised form 10 June 2021; Accepted 15 July 2021
Available online 23 July 2021

2014; Press, 2016; Wickham, 2014). A process of pre-harmonizing has been successful for some types of data in large community efforts. In some cases, the original investigators transform their data into a community-vetted, prescribed format using controlled terminology, such as Darwin Core-based contributions to the Global Biodiversity Information Facility (GBIF, 2021), or the observation model used by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI, Tarboton et al., 2008). In other cases, data collection and formatting efforts are coordinated from the start (e.g., Baldocchi et al., 2001; Duffy et al., 2019; Fraser et al., 2013; Leray and Knowlton, 2015; Mulholland et al., 2001; Stokstad, 2011). Prescribed formats are more easily achieved for some types of regular monitoring (e.g., sensor data), and the concept of Analysis-Ready data (ARD) is becoming prominent in the earth-observing field to reduce the burden of pre-processing on users (Dwyer et al., 2018). However, the idiosyncratic methods for collecting organismal data preclude most efforts to apply any single standard to spatial or taxonomic concepts, and standard data formats rarely find community acceptance because most cannot accurately capture complex environmental sampling conditions or other constraints particular to each research program (Kissling et al., 2018; Reichman et al., 2011). Furthermore, in many cases, incentives for the original researchers to transform their data are lacking. Ultimately, these barriers to synthesis of datasets inhibit collaboration and slow down potential scientific insights (Evans, 2016; Poisot et al., 2019).

Today, complex ecological datasets are becoming available from single locations where observations were collected consistently over long time periods. If combined appropriately, with the diversity in their sampling approaches overcome, these datasets become indispensable to understanding trends, testing ecological theory, and predicting changes in the numerous ecosystem services beneficial to society (Orth et al., 2020; Pereira et al., 2013). Research networks like the National Science Foundation's (NSF) Long Term Ecological Research (LTER) Network have met the expectation that their data are available in public repositories and permanently archived (Mayer, 2020; Servilla et al., 2016). These primary datasets are especially valuable and are increasingly being synthesized and reanalyzed to generate new knowledge (Collins et al., 2018; Dornelas et al., 2014; Record et al., 2021). This increased third-party use shows that datasets are now meeting some of the FAIR principles (Wilkinson et al., 2016), in that they are "Findable" and "Accessible". However, many would benefit from improvements to their interoperability and reusability, the "IR" of FAIR.

Here, we focus specifically on ecological community observation data and the collaboration among the Environmental Data Initiative (EDI) repository managers, data scientists from the National Ecological Observatory Network (NEON), and community ecologists from the LTER Network to recombine such data for reanalysis and improve their reusability. The need for this effort was prompted by community ecology synthesis working groups who noted that because pertinent datasets are formatted and described in a manner most appropriate to their unique original research objectives, they are not easily used in synthesis studies without major harmonization efforts. Multiple working groups typically use subsets of the same data independently and develop their own investigation-specific data cleaning, aggregation, and formatting procedures that do not translate across projects. This re-wrangling of datasets effectively duplicates large amounts of effort and impedes synthesis science insights, pointing to a need for a harmonization system for data collected at particular levels of biological organization (e.g., population, community, ecosystem; Record et al., 2021).

The harmonized format we present here is agnostic to the research question, adds specific metadata for improved discovery and reusability, and accommodates different types of measurements (e.g., count, percent cover, biomass), taxonomic resolutions, and nesting of sampling designs over space and time. Given use case requirements, the repository framework, and the need to emphasize the importance of sampling context, this model and workflow framework appeared to be the best compromise, and we look forward to feedback from users (e.g., https://github.com/EDIorg/ecocomDP/issues). Here, we report on the model itself, a library in the R language to assist with creation, access and exploration, metrics of the model's use to date, plus compatibility with a widely used biodiversity format, the Darwin Core Archive (DwC-A).

## 2. Methods

The project was carried out in three phases: *Design*, *Implementation*, and *Maintenance*. *Design* captures essential attributes of a science domain, considers past and present standardization efforts, and potential linkages to external authoritative systems to disambiguate meaning. The design phase leveraged the activities of science synthesis working groups and data management expertise to identify accurate and persistent data patterns. *Implementation* is accomplished through conversion of archived legacy data by data contributors or by EDI's data curation team, and is supported by data pattern documentation, best practices guides, and software libraries. *Maintenance* is achieved through programmatic workflows that automatically run when source data packages are updated.

### 2.1. Design

#### 2.1.1. Learning from existing approaches

We identified several ongoing or completed harmonization efforts using existing community observations and including datasets available from the EDI repository. All of these efforts used similar datasets from multiple sources, and all are one-time efforts with minimal plans for maintenance or updating harmonized data. In many cases, the resulting harmonized datasets were used to answer specific research questions and were then further changed or extended for additional uses. The abstract view of these datasets were potential models for general harmonization, and three in particular exemplify the need for a more broadly useable data model for observations – one which is also capable of structuring spatial information and taxonomy: 1) Popler, a database and R-libraries designed to analyze LTER population time series (Compagnoni et al., 2020); 2) CESTES, a global database for metacommunity ecology (Jeliazkov et al., 2020); and 3) BioTime, a global database of species abundances through time (Dornelas et al., 2014). In addition to the three research-focused models, we also considered the Darwin Core Archive (DwC-A) format used by the GBIF (Wieczorek et al., 2012). The GBIF system is arguably the largest aggregator of organismal occurrence and related data, holding over 1.5 billion records of species occurrences, taxonomic checklists, and sampling event or sample data from over 1500 institutions.

All three research-focused models implemented table structures and measurement types which do not accommodate the wide variety of raw data that capture complex environmental conditions during sampling and which are available in the original dataset. Only one (Popler) allows spatial nesting and taxon authority referencing. None of these databases accommodates references to external measurement dictionaries or ontologies. For all, access is somewhat limited by the choices of storage (i.e., Excel, or relational databases which require a custom interface or code). Temporal sampling is generally limited to observation dates, and CESTES includes text fields to describe nuances of temporal or other sampling. Compiled harmonization efforts such as these are highly valuable, as they represent considerable scientific knowledge and hours (possibly days) of thorough, manual checking and reformatting. Computing cannot supplant that scientific knowledge, but a comprehensive intermediate format can streamline some of the reformatting tasks.

GBIF's DwC-A came closest to meeting the requirements for broad reuse; these are self-contained datasets composed of text tables plus a file describing table organization. Table columns are labeled using the Darwin Core vocabulary (DwC) for indexing. A large fraction of GBIF records are simple organism occurrences, however DwC-A extensions allow for inclusion of other aspects such as contributor-defined measurements (e.g., abundance or cover), which are common for ecosystem studies of the type housed by EDI and data products published by NEON. The DwC also includes fields for external taxon references. Missing from the DwC-A were explicit site nesting and external measurement references (see Discussion). Interestingly, some of the structures created by scientists for their own synthesis can be strikingly similar to DwC-A tables (Walter et al., 2021) with features added (e.g., the aforementioned nested sampling sites).

### 2.1.2. Identifying requirements

Consistent with the goals to support a synthesis workflow that will reduce data preparation efforts for answering new research questions and minimize impact on data producers, we developed requirements based on three main considerations (see also discussion in Sholler et al., 2019): 1) the expectations of data contributors and the original data; 2) the repository framework; and 3) the needs of the data reusers. The scope is defined as ecological community data, in which observations are abundances of co-occurring groups of organisms in an area, as opposed to population or demographic data (where observations are made at the level of individuals within a species). We recognize that some original data will contain both types of information, and ideally, while the harmonized intermediate may not contain the original population-level information, the framework should make that original readily available. Our short name for a model for the flexible intermediate for ecological community data is "ecocomDP", for "ecological community data design pattern".

*2.1.2.1. Data contributors and the original data.* Original data are available in the EDI repository as text tables (usually ASCII) formatted to best suit the original research questions, with collection methods that are adapted to the environment and community of interest (e.g., aquatic, forest, grassland). In many cases the datasets are updated regularly. The data contributors (data managers or scientists) are intimately familiar with local conditions, which is vital to creating high-quality data packages. As mentioned above, there is no incentive for the data contributor to format their data in any other way, and so it was essential that the harmonization process did not interfere with a data contributor's formatting for their original research questions. The challenges presented by the data themselves included the large number of different parameters measured (e.g., number of individuals, cover, biomass, catch per unit effort), taxonomic resolution and consistency (e.g., family, genus, species), environmental or experimental conditions essential to interpretation (e.g., fertilization, harvest, simulated disturbance), the nesting of sampling units over space (e.g., site, transect, plot, subplot, depth) and time (e.g., date, season, year), plus changes to the sampling protocol over time (e.g., the addition of new sampling locations or changes in the taxonomic resolution of sampling).

Additionally, NEON publishes a variety of data products on its portal that provide biodiversity data on sentinel taxonomic groups from 81 field sites located across the United States (https://data.neonscience.org/). Many of these data products were designed with input from and for use by population and community ecologists (Thorpe et al., 2016; Utz et al., 2013). These products offer organismal data that can be mapped to the ecocomDP model, used in research, and derived data packages can then be archived in the EDI repository (e.g., Li et al., 2021).

*2.1.2.2. The repository framework.* In the EDI repository the granule is a "data package", composed primarily of a metadata record (Ecological Metadata Language, EML) and, one or more data entities (i.e., ASCII tables). The repository supports metadata and data immutability, revision control, DOI assignment and event subscriptions to track updates to data. Repository staff, although experienced data specialists, lack specific local knowledge for every dataset.

*2.1.2.3. The data users.* Aside from a standard data format and nomenclature, scientists attempting to use these existing data were mostly concerned with data discovery, i.e., the ability to identify data that best suited their needs in a repository. A few types of searches were common to all reuse (e.g., number of taxonomic units in study, duration of study and frequency of sampling, and the size and arrangement of sampling areas), and so needed to be supported. Those who are reformatting data to this model must understand the original data well, and so its associated code should include checks for certain features, like uniqueness and typing.

Our solution to these requirements is the development of a flexible domain-specific intermediate model in a lightweight, distributed workflow framework, in which data repositories handle some of the preparation work typically done by end users. The original data are not aggregated or otherwise changed, only normalized to a standard format that can be more readily accessed and used. This reformatting is accomplished by automated workflows which allow data products to be repeatedly synchronized when original data are updated. This process increases the value of the data by implementing standard quality checks and can provide feedback to contributors to inform them of aspects of data and metadata that are the most important during reuse, and of arrangement or presentation choices that function well.

## 2.2. Implementation

During the implementation phase, pertinent datasets in the EDI and NEON repositories were identified. For each EDI dataset, an R script was developed to convert the data into the ecocomDP model. This effort incrementally led to tuning of the model itself and associated documentation. It also served to outline necessary functions for building data packages and accessing NEON data. Lastly, to test both the data format and the entire workflow, we used ecocomDP formatted data to generate DwC-A for submission to GBIF. This last step has the added benefit of making EDI holdings available for GBIF users.

Fig. 1 depicts the general workflow which was implemented and will be followed for updates. The "level" designations and terminology are adapted from NASA's Earth Observing System Data and Information System (EOSDIS) (Price et al., 1994) with L0 being the original data; L1 is the same data transformed to the ecocomDP model, and made available as a data package in the EDI repository. L2 has been further transformed or aggregated as needed for a particular synthesis research question or other use (such as a DwC-A).

## 2.3. Maintenance

The maintenance phase focuses on developing robust R scripts for continued conversion when source data (L0) are updated and converting new datasets as they are submitted to the EDI repository. Maintenance of the R package includes adaptations for the NEON endpoints as these evolve. The EDI infrastructure supports the execution of external workflows through its API and event notification service to automate routine data management tasks. Upload of an L0 revision triggers execution of its conversion script. The system is ideal for a series of data packages, as it simplifies and accelerates creation of continuously updated synthetic data packages (Servilla et al., 2016).

# 3. Results

## 3.1. The ecocomDP data model

The model (Fig. 2) is composed of eight related data tables in an extended star schema (Seyed-Abbassi and Madesi, 2015) and implements database-style principles of foreign keys and normalization, along with attribute/value style tables to accommodate a wide range of measurements. Three data tables are required: the central "observation" table and two supporting dimensional tables, "sampling_location", and

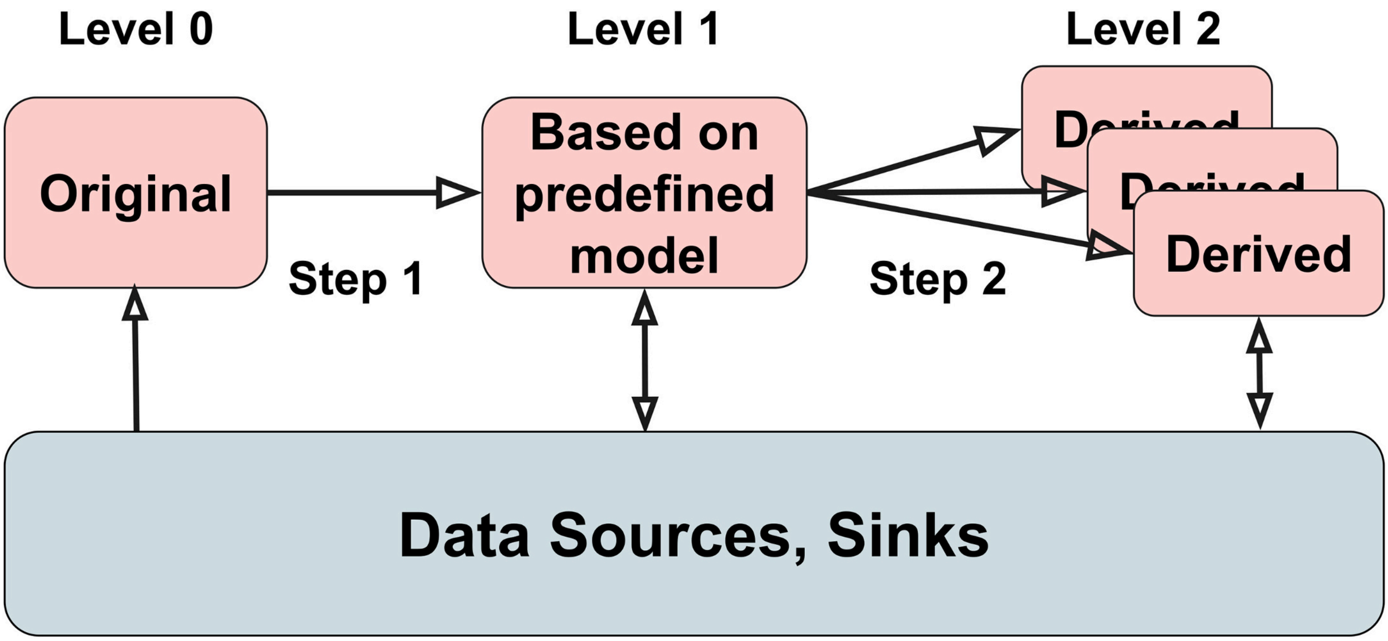


**Fig. 1.** Level 0 (L0) are incoming, original data, ideally, already archived in the repository with complete metadata and contributed by those close to the research. Level 1 (L1) data packages (also in the repository) are formatted according to a predefined model, in this case, ecocomDP. Researchers 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).

"taxon". The "dataset_summary" is automatically created and populated based on the observations. The three primary tables are each extended with an optional table for ancillary information to accommodate additional measurements important to understand and use specific sampling conditions for analysis. The optional eighth table maps variables to external dictionaries.

#### 3.1.1. *Observations*

The central "fact" table holds the actual ecological community observations (Fig. 2, e.g., abundances or densities of a taxon).

#### 3.1.2. *Locations*

The nesting of sampling locations (e.g., plots within transects within areas, or depths or heights of a profile) is accomplished using a self-referencing table, in which a location may have a 'parent' which is itself a sampling location in the same table. This mechanism allows observations to be associated with a location at any level, and observations can be aggregated under groups of locations.

#### 3.1.3. *Taxonomy*

The taxonomy table does not attempt to describe all aspects of a taxon, but rather holds basic information such as name and rank (e.g., family, genus, species), with the option to refer to a taxonomic name authority system. Although a taxonomic name may be reused in different kingdoms and a hierarchy required for full understanding, the model deliberately does not encode taxonomic hierarchies, as these are somewhat fluid and no single system applies to all organisms. Instead, that information can be held by the authority system, and accessed with readily available software tools, or it can be recorded in the taxon_ancillary table.

#### 3.1.4. *Summary table*

A one-row table summarizes information in the Observation, Location, and Taxonomy tables. It represents the information most frequently needed by scientists as they evaluate a dataset for use, mainly to understand the taxonomic, temporal, and spatial coverage.

#### 3.1.5. *Ancillary tables*

Each primary table has an optional table for additional information. Also designed as attribute/value, these ancillary tables provide a place for environmental conditions (e.g., air temperature, observation uncertainties), organism characteristics, (e.g., biomass, traits, morphotype, phylogenetic information), or experimental conditions (e.g., fertilization). Date fields are included for taxon_ancillary and location_ancillary as these may have been recorded a different times than the primary observation. The observation_ancillary table might contain specific sampling-event-data, such as volume cleared by a plankton tow or single depth (when not part of a profile). These are data typically included with the community observation data to ensure that data users are aware of conditions and can judiciously subset and aggregate original observations.

#### 3.1.6. *Accommodating measurement term disambiguation*

An optional "variable_mapping" table allows unambiguous term definition using external vocabularies and ontologies by documenting the system used and a unique identifier for the term (i.e., a URI or URL). It is intended for the content of fields titled 'variable_name' in the observation and optional ancillary tables.

### 3.2. *Supporting code*

We developed an open-source code library in the R statistical language to support common tasks for creating, checking and using ecocomDP data packages (Smith and Sokol, 2021). To assist with conversion to ecocomDP from EML-described data packages, R functions are available to harvest EML metadata from the L0 dataset preserving essential high-level elements (e.g., abstract, methods and personnel), with additional text and EML elements to clarify that this (L1) is a derived data product: a provenance link to the L0 dataset, additional abstract and title text, and keywords (e.g., "ecocomDP"). L0 variable names and descriptions are transferred to coded value lists in L1 EML. To promote discovery, some ecocomDP table content is elevated to metadata, such as full taxonomic hierarchies including common names and external identifiers, and EML annotations created from the

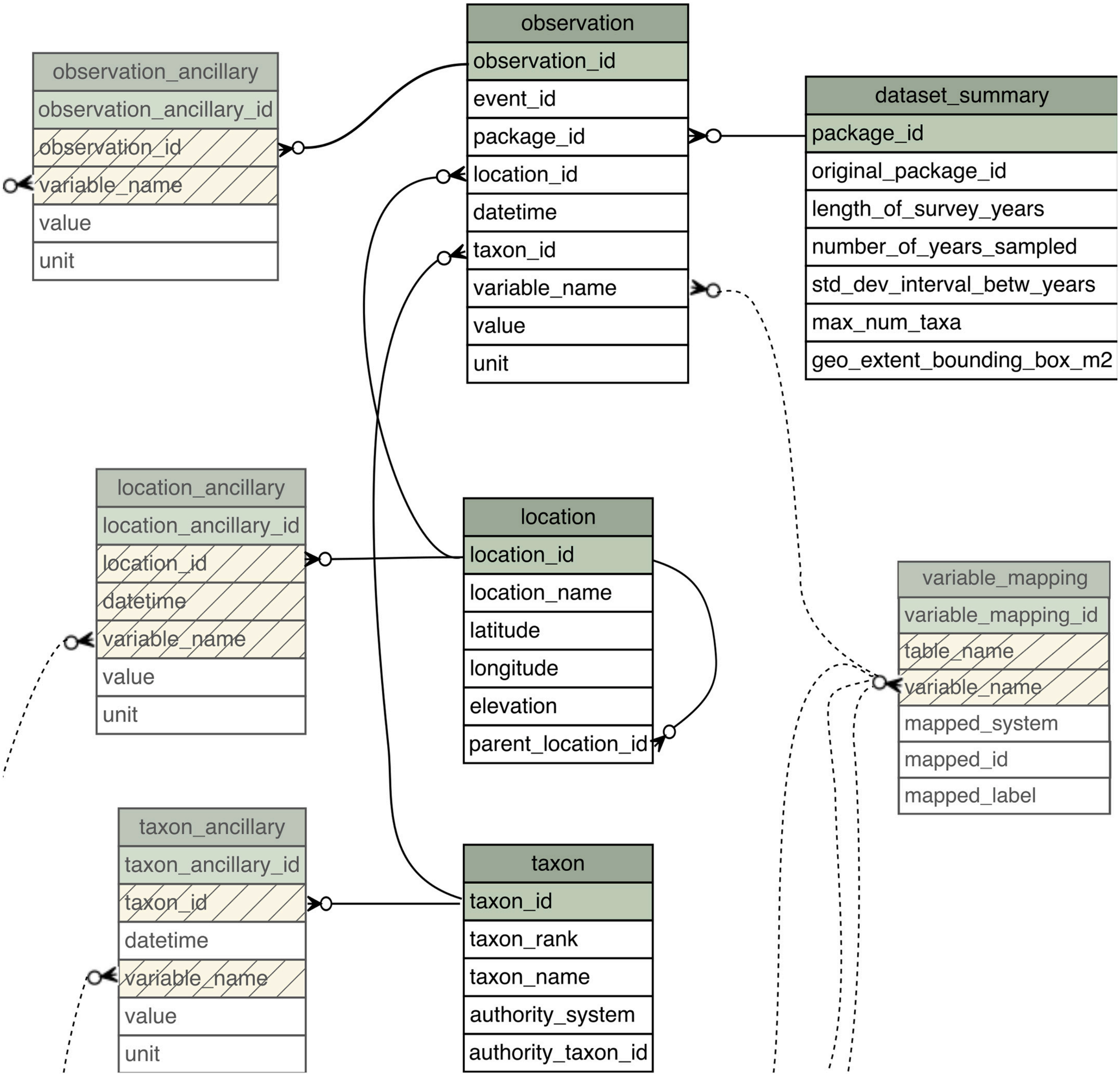


**Fig. 2.** The ecocomDP model shown with relational database notation for foreign keys and relationships (e.g, lines ending in crows-foot indicate 1:many relationships). Semi-transparent tables are optional. Medium green fields in each table are the primary key. Yellow/hashed fields are a combined unique constraint. IDs (suffixed, "_id"), must be unique within a table, as in an relational database. Full documentation (e.g, optional fields and definitions) can be found in the Git repository (EDI, n.d.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variable_mappings table. The R library also supports quality control to ensure that tables are model-compliant, confirming presence of required fields, referential integrity between tables, and uniqueness of identifiers. Taxon IDs are added with the taxize R library (Chamberlain et al., 2020).

The ecocomDP R library provides functions to search data and metadata on free text, taxonomic names, geographic area, and summary features (from the dataset_summary table, Fig. 2), which is improved over typical repository searches on metadata alone. Analysis workflows are supported through functionality for programmatically accessing and reading the data and metadata; merging datasets; transposing ecocomDP tables into the "wide" format (e.g., each column representing a taxon or variable) preferred by many scientists; and for creating plots of basic features to evaluate fitness for use (see below). As we have already stated, preparing data for analysis can still be complex, and these tools will not replace ecological understanding of fitness for use of data in a particular analysis. However, they will help streamline the process considerably.

### *3.3. Using the ecocomDP format*

The R library described above was developed and tested as we processed original, incoming data through the Fig. 1 workflow, first converting them to the ecocomDP model (L1; Fig. 1, Step 1), followed by a) plotting general characteristics as might be required by synthesis and b) conversion to publication ready DwC-A (an example of L2). Those processes and summary metrics from conversions are detailed here. As incoming datasets are nearly always unique, the conversion to the ecocomDP format (L1) requires an understanding of the study design, measurement methods and data types, with the R library helping to ensure full understanding, and appropriate use and accelerating the technical steps. Because all L1 are a standard format, further processing can be streamlined, and often automated.

### *3.3.1. Converting original data to ecocomDP (L0 to L1)*

To date, we have created 70 ecocomDP data packages from EDI holdings of LTER, Long Term Research in Environmental Biology (LTREB), and other projects. Our approach to conversion of these original (L0 datasets) is to assemble each package's data into a single wide table, which helps maintain referential integrity in the derived tables. Issues arising at this step are best resolved in collaboration with the original data creators and may provide valuable feedback to them. The next step is to extract data from the L0-wide table for the core ecocomDP tables (i.e., observation, taxon, location; Fig. 2) followed by the optional ancillary tables. The ecocomDP R library supports common steps for scripting the entire process, including programmatic reading of the L0 package. We recommend scripting this entire step for two reasons: the script serves as documentation of the process, and if the L0 data package is updated (e.g., new data added), subsequent conversions can be automated.

When the original data format is well controlled, reformatting to the ecocomDP model is more straightforward. NEON exposes its corpus of datasets of organism data for integration with EDI's holdings, using code created by NEON with scientists from the NEON Science Summit Meeting (Boulder, CO, 2019) (Li et al., 2021). R functions pull data from the NEON share point using the neonUtilities R library and convert it from a NEON data product to the ecocomDP data pattern. As of this writing, functions are available in the ecocomDP R library to deliver data for NEON terrestrial organisms (breeding land birds, DP1.10003.001; ground beetles, DP1.10022.001; herptile bycatch from ground beetle sampling, DP1.10022.001; small mammals, DP1.10072.001; mosquitoes, DP1.10043.001; terrestrial plants, DP1.10058.001; ticks, DP1.10093.001; tick pathogens, DP1.10092.001) and for aquatic organisms (fish, DP1.20107.001; macroinvertebrates, DP1.20120.001; microalgae, DP1.20166.001; zooplankton, DP1.20219.001) at all sites where NEON routinely collects those data.

As NEON data products are continent-wide, these were divided into individual field sites for analysis to make them spatially compatible with EDI holdings. For both NEON and EDI data, summary information, identifiers and DOIs if applicable can be found in the dataset, O'Brien et al. (2021). Spatial, temporal, and taxonomic coverage for a total of 530 NEON and EDI datasets are shown in Fig. 3, comprising over nine million observations. The NEON data are broken out by sites (83 total sites) as that unit was more similar in structure to the data packages available from EDI, which come from site-based research groups such as the LTER Network. Data in harmonized format clearly illustrate the differences between the data collection strategies of NEON and the EDI holdings from individual place-based sampling programs. NEON's targeted biological collections focus on nine groups of species (by taxonomic or other attributes) over relatively narrow spatial extents within sites (but a large spatial extent among sites), and over shorter, evenly-

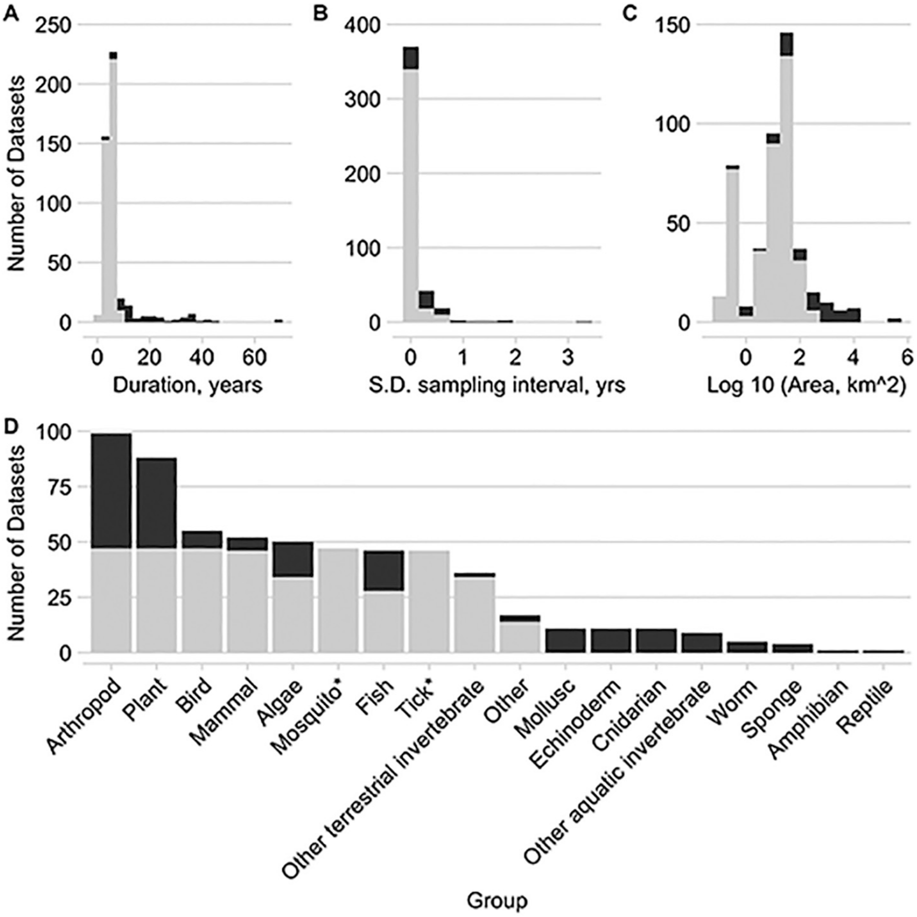


**Fig. 3.** Temporal, spatial and taxonomic coverage of datasets available in the ecocomDP model. Data source: Black, EDI; Gray, NEON. A) Temporal coverage (years), B) Temporal evenness (years), C) Spatial extent, D) group. An asterisk indicates that two groups (Tick, Mosquito) are specifically targeted by NEON. When these taxa occur in EDI datasets, they are plotted here with Arthropods.

spaced time periods (collections began in 2013 with full operations in 2019). Coverage plotted from EDI data holdings, on the other hand, shows a wide diversity for all three coverage elements and reflects the diversity of research programs. Durations range from a few years to over six decades, with somewhat less even sampling, a broader spatial extent (up to $10^5$ km$^2$), and many general taxonomic groups represented.

### 3.3.2. *Working with ecocomDP formatted (L1) datasets*

The principles of a central observation table linked to additional information and the attribute/value pattern that underlies the ecocomDP model are common approaches for managing heterogeneous data due to their flexibility and storage efficiency (Wieczorek et al., 2012). We used the formatted data to demonstrate two outcomes: first, the ease of creating common plots for scientific evaluation, and second, a mechanism to create DwC-A for GBIF.

As with the coverage plots (Fig. 3), a common format enables other common plots to be created. The ecocomDP R library supports plotting of features commonly requested by scientists to evaluate a dataset's suitability for use. Fig. 4 shows four aspects: number of taxa over time, spatio-temporal sampling effort, species accumulation, and species shared among sites. These examples, plotted from L1 data represent features of interest to synthesis working groups and are based on their input (Jarzyna et al., 2021; Record et al., 2021; Walter et al., 2021). Community ecologists often use data on taxon presence or abundance to generate evidence that quantifies the strength of species interactions such as competition, predation, or mutualism, or responses to shared environmental conditions. For example, Record et al. (2021) used the L1 output to explore spatial and temporal representativeness of several LTER datasets to assess the suitability of LTER community datasets for addressing questions of how spatiotemporal scales influence insights from metacommunity analyses. Likewise, Jarzyna et al. (2021) used the L1 output of NEON data to explore temporal dynamics in animal communities at a continental scale. Walter et al. (2021) synthesized the spatial synchrony of biodiversity across 20 marine and terrestrial communities. The ability to quickly create the common plots shown in Fig. 4 for many datasets were instrumental in streamlining the data-discovery phase of each of these syntheses.

In addition to supporting reuse of community observation data by synthesis science, a goal for this harmonization effort is to contribute these datasets to the holdings of the Global Biodiversity Information Facility (GBIF) to increase the data's discovery and use. Although the ecocomDP model is more extensive than the DwC-A, their similarities make a scripted process straightforward. Both the DwC-A and ecocomDP models are star schemas with attribute/value tables and both use EML for metadata. Information loss is minimized by mapping to DwC-A's Event Core layout (GBIF, 2021). Our approach makes use of ecocomDP R functions for manipulating datasets, followed by mapping to the DwC-A terms and adding required metadata elements. Several types of external identifiers are included in the DwC-A tables. For taxa, we include ids (DC: taxonID) with named authority (DC: nameAccordingTo) and Life Science Identifiers (LSIDS) in the DC scientificNameID field. We also make use of the recently added EML annotation field (Jones et al., 2019) to include measurement URIs in the DwC-A extension field measurementTypeID.

With the conversion from original (L0) data to ecocomDP (L1) formatted data to DwC-A (L2) data fully automated, updating long-term observational datasets is simplified. As of this writing, we are working with GBIF on the technical aspects of the contribution mechanism. In the interim, all DwC-A packages are in the EDI data portal via the keyword "Darwin Core Archive". Researchers will soon have several options for accessing these data in addition to the original dataset: the ecocomDP-formatted and the archived DwC-A packages both archived at EDI, and by querying values through GBIF systems.

## 4. Discussion

Decades of harmonizing data from diverse studies and developing community data standards at multiple scales indicate that a substantial upfront cost is incurred. These laborious efforts must be justified by benefits such as importance to meta-analyses, reduced expense of obtaining and preparing them for analysis, or even commercial value. Further, it appears that harmonization efforts generally lead to a certain loss of information, which can be acceptable during analysis if balanced by sufficient volume (e.g., Pollet et al., 2015). As a result, highly complex, multidimensional data have largely eluded harmonization. Ecological community observations, although irreplaceable and highly valued for understanding environmental change (LTERnet.edu n.d), have highly-variable sampling methods and high dimensionality that

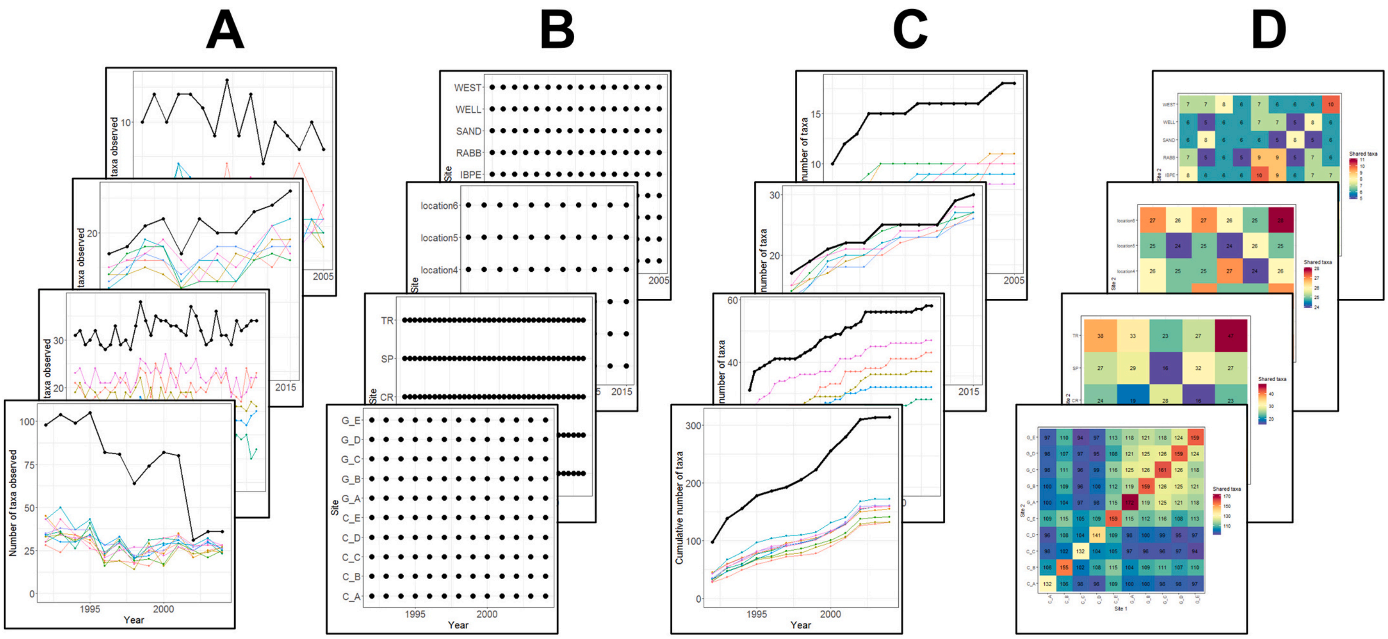


**Fig. 4.** Plots from four L1 datasets. (A) number of unique taxa (y-axis) observed over time (x-axis), (B) sampling effort over time (x-axis) and space (y-axis), (C) species accumulation curves (y-axis) over time (x-axis), and (D) matrix of species co-occurring among sites (site 1 on x-axis and site 2 on y-axis).

continue to make synthesis across studies difficult (Welti et al., 2021). A level of pre-harmonization is essential if the community is to avoid each synthesis group expending significant effort repeatedly wrangling data into similar formats, and to promote more rapid and reproducible synthesis efforts (Record et al., 2021).

Given these experiences, requirements, and use cases, our new data model minimizes information loss while meeting most of the needs of meta-analysis, and uses a workflow system that also accounts for regular updates to the datasets. The reformatted data (ecocomDP format) are maintained as independent packages in the EDI repository to take advantage of its general functionality of search and access, hence avoiding another database 'silo'. Further, specific discoverability is improved by the addition of standardized metadata to aid the process of selecting relevant datasets. Any synthesis effort will still have the significant step of determining if a dataset is fit for a particular analysis, which is typically performed by examining the sampling methods, constraints, and other facets of data collection. That task can be further assisted by disambiguating semantics through linkages to external dictionaries, which is accommodated in the ecocomDP data model as well as the EML metadata standard. Li et al. (2021) details the decisions made while converting NEON data to ecocomDP. Some of the checking available in our R-package is a result of that, however additional dependencies or checks may become evident which help ensure that scientists fully understand the data as they convert it into the ecocomDP format.

Although extensive reusable R programming functionality was developed, the conversion from original data formats (L0) to ecocomDP format (L1) still requires a moderate investment in time and some ecological understanding for every new dataset—a significant task taken on primarily by the repository, EDI. Future reuse of these data will determine the value of such a reformatting service and the likelihood of its continuation. An advantage of the workflow system is that after the initial effort, the scripts generating ecocomDP data packages from the original data can be fully automated and repeated when the original data are updated. The generation of downstream data products can also be automated, and our creation of DwC-A for submission to GBIF serves as a model for generating submissions to other systems, such as Popler, CESTES, BioTIME or VegBank (Peet et al., 2012). In addition to supporting short-term synthesis research, we envision these important data supporting the needs of ecological forecasting studies (e.g., Dietze et al., 2018) and being used to calculate indices for Essential Biodiversity Variables (EBV, (Pereira et al., 2013; GEO-BON, 2013), the community-managed state variables that stand between primary observations, or even for higher-level indicators such as the Ocean Health Index, (Halpern et al., 2012, 2015; Schmeller et al., 2015).

The flexible attribute/value data format used for ecocomDP has been widely used in other data harmonization approaches (e.g., Tarboton et al., 2008; Wieczorek et al., 2012). It saves space and allows an unlimited number of attributes, hence accommodating any type of measurement. However, description and control of aspects such as data typing, precision, or text definitions are not built in, and as compared to the detailed data table descriptions common in the original data packages, may result in some metadata loss. The ecocomDP project mitigates such losses by retaining as much metadata as possible, quality checking, and by implementing a workflow system that includes a provenance trace in derived data (L1, L2; Fig. 1) so that original data can be accessed if necessary.

The semantic parity between ecocomDP and the DwC-A model is strong, especially for concepts like Observation and Taxon. The GBIF and Darwin Core systems work quite well for observations of individuals but less well for measures of abundance; the ecocomDP model helps fill that gap. The functionality of ecocomDP's ancillary tables is aligned with ExtendedMeasurementOrFact, and together these features helped to streamline our conversion to DwC-A. Although that conversion was relatively straightforward, there are significant differences between the two formats. First, the DwC vocabulary and GBIF model does not explicitly support the kind of site nesting needed to understand a sampling design. The Event class (which includes locations) can be leveraged for this use (De Pooter et al., 2017), although examples and recommendations are not well-established in the community. Therefore, ecocomDP explicitly includes a site-nesting feature, similar to other models used by scientists (i.e., Popler, Compagnoni et al., 2020). Our conversion scripts can be adapted in the future as the use of the DwC-based models evolves. Secondly, inclusion of external dictionary references for measurements is not currently an established part of the DwC vocabulary (which determine column headings for DwC-A). Our L2 DwC-A already includes the proposed extension for measurementID (to hold URIs in external measurement dictionaries) and will serve as an example as adoption of this extension increases. Those differences, and the ease with which our ecocomDP datasets can be converted to DwC-A makes the ecocomDP intermediate valuable both for detailed scientific syntheses and large-scale querying by aggregators like GBIF.

The use of ecocomDP to promote discovery, reusability, and integration of data is an exciting step towards harmonization of data across coordinated research networks, which advances collating in-situ ecological community observation data at global extents to support broad concepts such as EBVs. This EDI-NEON collaboration also reveals the value of synergies between networks by integrating the deep long-term and place-based knowledge of the LTER Network with the broad spatial coverage of the NEON Observatory. Just as harmonization of data helps synthesis scientists avoid "reinventing the wheel" for each research project, collaboration among groups such as NEON, LTER, and EDI promotes communication between repository staff and scientists to share insights and pitfalls about data. Furthermore, although NEON data are extremely well documented and encapsulate standardized collection protocols, the level of detail surrounding slight nuances in data collection over time (e.g., reductions in sampling events) or abbreviations used (e.g., "sp." and "spp.") may elude users. The oversight of data wrangling in collaboration with NEON staff for the ecocomDP model assures users that these idiosyncrasies have been considered. End users will still need to recognize that the ecocomDP data are intended to be used for community ecology analyses rather than for demographic analyses, although the original data may contain that information. For instance, to access NEON's small mammal mark-recapture information (e.g., to estimate occupancy for population models) users would need to return to the the original data product.

## 5. Conclusion

Many important primary data are ongoing research-grade time series, and access to these trusted, up-to-date data sources is highly desired by synthesis scientists, managers, and policy and decision makers, yet easy access is seldom realized. Data harmonization is not a new idea. But typically, harmonization projects for organismal data are designed for specific research questions or types of queries, which tend to drive data preparation decisions. Unfortunately, those formatting or aggregation choices often reduce the potential for other types of use.

Our workflow-based model makes both the original data and harmonized version easy to discover and access, and takes advantage of existing repository functionality. Furthermore, heterogeneous data become available in a manner consistent and interoperable with current and emerging trends in other biological fields. The harmonized intermediate has basic formatting applied, and accommodates standardized measurement semantics and taxonomy. The use of event subscriptions to track their updates and rerun processing code is a transformative activity, and provides a template for a process that can be reused in other scientific domains.

## Declaration of Competing Interest

None.

## Acknowledgements

This work was supported by the National Science Foundation grant numbers 1931143 to M. Servilla, 1931174, to C. Gries, 1926568 to S. Record, and 1545288 and 1929393 to F. Davis.

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

This material is based in part upon work supported by the National Science Foundation through the Long Term Ecological Research (LTER) Program. LTER's Network Office is based at the National Center for Ecological Analysis and Synthesis, UC Santa Barbara.

We thank Dr. Nathan Wisnoski (University of Wyoming) for the original code used in Fig. 4D.

## References

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Wofsy, S., 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82 (11), 2415–2434. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2.

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, I., Baumgartner, J., O'Donnell, J., Oksanen, J., Tzovaras, B.G., Marchand, P., Tran, V., Salmon, M., Li, G., Grenié, M., 2020. taxize: Taxonomic Information from Around the Web. R Package Version 0.9.98. https://github.com/ropensci/taxize.

Collins, S.L., Avolio, M.L., Gries, C., Hallett, L.M., Koerner, S.E., La Pierre, K.J., Rypel, A.L., Sokol, E.R., Fey, S.B., Flynn, D.F.B., Jones, S.K., Ladwig, L.M., Ripplinger, J., Jones, M.B., 2018. Temporal heterogeneity increases with spatial heterogeneity in ecological communities. Ecology 99 (4), 858–865. https://doi.org/10.1002/ecy.2154.

Compagnoni, A., Bibian, A.J., Ochocki, B.M., Levin, S., Zhu, K., Miller, T.E.X., 2020. Popler: an r package for extraction and synthesis of population time series from the long-term ecological research (LTER) network. Methods Ecol. Evol. 11 (2), 258–264. https://doi.org/10.1111/2041-210X.13319.

De Pooter, D., Appeltans, W., Bailly, N., Bristol, S., Deneudt, K., Eliezer, M., Fujioka, E., Giorgetti, A., Goldstein, P., Lewis, M., Lipizer, M., Mackay, K., Marin, M., Moncoiffé, G., Nikolopoulou, S., Provoost, P., Rauch, S., Roubicek, A., Torres, C., van de Putte, A., Vandepitte, L., Vanhoorne, B., Vinci, M., Wambiji, N., Watts, D., Klein Salas, E., Hernandez, F., 2017. Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiv. Data J. 5, e10989 https://doi.org/10.3897/BDJ.5.e10989.

Dietze, M.C., Fox, A., Beck-Johnson, L.M., Betancourt, J.L., Hooten, M.B., Jarnevich, C.S., Keitt, T.H., Kenney, M.A., Laney, C.M., Larsen, L.G., Loescher, H.W., Lunch, C.K., Pijanowski, B.C., Randerson, J.T., Read, E.K., Tredennick, A.T., Vargas, R., Weathers, K.C., White, E.P., 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. Proc. Natl. Acad. Sci. 115 (7) https://doi.org/10.1073/pnas.1710231115.

Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., Magurran, A.E., 2014. Assemblage time series reveal biodiversity change but not systematic loss. Science 344 (6181), 296. https://doi.org/10.1126/science.1248484.

Duffy, J.E., Benedetti-Cecchi, L., Trinanes, J., Muller-Karger, F.E., Ambo-Rappe, R., Boström, C., Buschmann, A.H., Byrnes, J., Coles, R.G., Creed, J., Cullen-Unsworth, L.C., Diaz-Pulido, G., Duarte, C.M., Edgar, G.J., Fortes, M., Goni, G., Hu, C., Huang, X., Hurd, C.L., Yaakub, S.M., 2019. Toward a coordinated global observing system for seagrasses and marine macroalgae. Front. Mar. Sci. 6. https://doi.org/10.3389/fmars.2019.00317.

Dwyer, J.L., Roy, D.P., Sauer, B., Jenkerson, C.B., Zhang, H.K., Lymburner, L., 2018. Analysis ready data: enabling analysis of the Landsat archive. Remote Sens. 10 (1363-undefined).

EDI, n.d. EDIorg/ecocomDP. GitHub.com. https://github.com/EDIorg/ecocomDP (last accessed 27 July 2021).

Evans, S.R., 2016. Gauging the purported costs of public data archiving for long-term population studies. PLoS Biol. 14 (4), e1002432 https://doi.org/10.1371/journal.pbio.1002432.

Fraser, L.H., Al Henry, H., Carlyle, C.N., White, S.R., Beierkuhnlein, C., Cahill, J.F., Casper, B.B., Cleland, E., Collins, S.L., Dukes, J.S., Knapp, A.K., Lind, E., Long, R., Luo, Y., Reich, P.B., Smith, M.D., Sternberg, M., Turkington, R., 2013. Coordinated distributed experiments: An emerging tool for testing global hypotheses in ecology and environmental science. In: Frontiers in Ecology and the Environment, 11. John Wiley & Sons, Ltd, pp. 147–155. https://doi.org/10.1890/110279. Issue 3.

GEO-BON, 2013. "What are EBVs?". Group on Earth Observations. https://geobon.org/ebvs/what-are-ebvs/ (last accessed 2021-07- 27).

GBIF, 2021. Introduction to Sampling Event Data. https://www.gbif.org/sampling-event-data (n.d.).

Halpern, B.S., Longo, C., Hardy, D., McLeod, K.L., Samhouri, J.F., Katona, S.K., Kleisner, K., Lester, S.E., O'Leary, J., Ranelletti, M., Rosenberg, A.A., Scarborough, C., Selig, E.R., Best, B.D., Brumbaugh, D.R., Chapin, F.S., Crowder, L.B., Daly, K.L., Doney, S.C., Zeller, D., 2012. An index to assess the health and benefits of the global ocean. Nature 488 (7413), 615–620. https://doi.org/10.1038/nature11397.

Halpern, B.S., Longo, C., Lowndes, J.S.S., Best, B.D., Frazier, M., Katona, S.K., Kleisner, K.M., Rosenberg, A.A., Scarborough, C., Selig, E.R., 2015. Patterns and emerging trends in Global Ocean health. PLoS One 10 (3). https://doi.org/10.1371/journal.pone.0117863.

Heffernan, J.B., Soranno, P.A., Angilletta Jr., M.J., Buckley, L.B., Gruner, D.S., Keitt, T.H., Kellner, J.R., Kominoski, J.S., Rocha, A.V., Xiao, J., Harms, T.K., Goring, S.J., Koenig, L.E., McDowell, W.H., Powell, H., Richardson, A.D., Stow, C.A., Vargas, R., Weathers, K.C., 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. Front. Ecol. Environ. 12 (1), 5–14. https://doi.org/10.1890/130017.

Jarzyna, M.A., Norman, K.E.A., LaMontagne, J.M., Helmus, M.R., Li, D., Parker, S.M., Rocha, M.P., Record, S., Sokol, E.R., Zarnetske, P.L., Surasinghe, T., 2021. Ecosystem stability is related to animal community dynamics at a continental scale. Ecosphere. (in review).

Jeliazkov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L., Barsoum, N., Bartonova, A., Belskaya, E., Bonada, N., Brind'Amour, A., Carvalho, R., Castro, H., Chmura, D., Choler, P., Chong-Seng, K., Cleary, D., Cormont, A., Cornwell, W., Chase, J.M., 2020. A global database for metacommunity ecology, integrating species, traits, environment and space. Sci. Data 7 (1). https://doi.org/10.1038/s41597-019-0344-7.

Jones, M.B., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Earl, S., Chong, S., 2019. Ecological Metadata Language version 2.2.0. KNB Data Repository. https://doi.org/10.5063/F11834T2.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J.-B., Obst, M., Santamaria, M., Skidmore, A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biol. Rev. 93 (1) https://doi.org/10.1111/brv.12359.

Leray, M., Knowlton, N., 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc. Natl. Acad. Sci. U. S. A. 112 (7), 2076–2081. https://doi.org/10.1073/pnas.1424997112.

Levy, O., Ball, B.A., Bond-Lamberty, B., Cheruvelil, K.S., Finley, A.O., Lottig, N.R., Punyasena, S.W., Xiao, J., Zhou, J., Buckley, L.B., Filstrup, C.T., Keitt, T.H., Kellner, J.R., Knapp, A.K., Richardson, A.D., Tcheng, D., Toomey, M., Vargas, R., Voordeckers, J.W., Williams, J.W., 2014. Approaches to advance scientific understanding of macrosystems ecology. Front. Ecol. Environ. 12 (1), 15–23. https://doi.org/10.1890/130019.

Li, D., Record, S., Sokol, E.R., Bitters, M.E., Chen, M.Y., Chung, A., Helmus, M., Jaimes, R., Jansen, L., Jarzyna, M.A., Just, M.G., LaMontagne, J.M., Melbourne, B., Moss, W., Norman, K., Parker, S., Robinson, N., Seyednasrollah, B., Smith, C., Zarnetske, P.L., 2021. Tidy NEON data for biodiversity research. Ecosphere in prep.

Lohr, S., 2014, August 17. For Big-Data ScienIsts, 'Janitor Work' Is Key Hurdle to Insights. The New York Times. https://nyi.ms/1t8IzfE.

Mayer, A., 2020. Long term ecological research network celebrates 40 years of discovery. BioScience 70 (11). https://doi.org/10.1093/biosci/biaa098.

Mulholland, P.J., Fellows, C.S., Tank, J.L., Grimm, N.B., Webster, J.R., Hamilton, S.K., Martí, E., Ashkenas, L., Bowden, W.B., Dodds, W.K., Mcdowell, W.H., Paul, M.J., Peterson, B.J., 2001. Inter-biome comparison of factors controlling stream metabolism. Freshw. Biol. 46 (11), 1503–1517. https://doi.org/10.1046/j.1365-2427.2001.00773.x.

O'Brien, M., Smith, C.A., Sokol, E.R., Gries, C., Lany, N., Record, S., Castorani, M.C., 2021. EDI and NEON dataset descriptions and coverage to support the paper "ecocomDP: a flexible data design pattern for ecological community survey data" ver 1. Environ. Data Initiative. https://doi.org/10.6073/pasta/dc9ac7435f98c0a0c3a583f8a695899f ([dataset] Accessed 2021-06-01).

Orth, R.J., Lefcheck, J.S., McGlathery, K.S., Aoki, L., Luckenbach, M.W., Moore, K.A., Oreska, M.P.J., Snyder, R., Wilcox, D.J., Lusk, B., 2020. Restoration of seagrass habitat leads to rapid recovery of coastal ecosystem services. Sci. Adv. 6 (41) https://doi.org/10.1126/sciadv.abc6434 eabc6434.

Peet, R.K., Lee, M.T., Jennings, M.D., Faber-Langendoen, D., 2012. VegBank: a permanent, open-access archive for vegetation plot data. Biodiv. Ecol. 4, 233–241.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Wegmann, M., 2013. Essential biodiversity variables. In: Science, 339. American Association for the Advancement of Science, pp. 277–278. https://doi.org/10.1126/science.1229931. Issue 6117.

Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D., Peres-Neto, P., 2019. Ecological data should not be so hard to find and reuse. In: Trends in Ecology and Evolution, 34. Elsevier Ltd, pp. 494–496. https://doi.org/10.1016/j.tree.2019.04.005. Issue 6.

Pollet, T.V., Stulp, G., Henzi, S.P., Barrett, L., 2015. Taking the aggravation out of data aggregation: a conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. Am. J. Pimatol. 77, 727–740. https://doi.org/10.1002/ajp.22405.

Press, G., 2016. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes, March 23 2016. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says (acessed 2021-05-25).

Price, R.D., King, M.D., Dalton, J.T., Pedelty, K.S., Ardanuy, P.E., Hobish, M.K., 1994. Earth science data for all: EOS and the EOS data and information system. Photogramm. Eng. Remote. Sens. 60, 277–285.

Record, S., Voelker, N.M., Zarnetske, P.L., Wisnoski, N.I., Tonkin, J.D., Swan, C.M., Marazzi, L., Lany, N., Lamy, T., Compagnoni, A., Castorani, M.C.N., Andrade, R., R., & Sokol, E. R., 2021. Novel insights to be gained from applying metacommunity theory to long-term, spatially replicated biodiversity data. Front. Ecol. Evol. https://doi.org/10.3389/fevo.2020.612794.

Reichman, O.J., Jones, M.B., Schildhauer, M.P., 2011. Challenges and opportunities of open data in ecology. Science 331 (6018), 703. https://doi.org/10.1126/science.1197962.

Schmeller, D.S., Julliard, R., Bellingham, P.J., Böhm, M., Brummitt, N., Chiarucci, A., Couvet, D., Elmendorf, S., Forsyth, D.M., Moreno, J.G., Gregory, R.D., Magnusson, W.E., Martin, L.J., McGeoch, M.A., Mihoub, J.-B., Pereira, H.M., Proença, V., van Swaay, C.A.M., Yahara, T., Belnap, J., 2015. Towards a global terrestrial species monitoring program. J. Nat. Conserv. 25 https://doi.org/10.1016/j.jnc.2015.03.003.

Servilla, M., Brunt, J., Costa, D., McGann, J., Waide, R., 2016. The contribution and reuse of LTER data in the provenance aware synthesis tracking architecture (PASTA) data repository. Ecol. Inform. 36, 247–258. https://doi.org/10.1016/j.ecoinf.2016.07.003.

Seyed-Abbassi, B., Madesi, V., 2015. Data Warehouse Design Using Row and Column Data Distribution. Proceedings of the Int'l Conf. Information and Knowledge Engineering.

Sholler, D., Ram, K., Boettiger, C., Katz, D.S., 2019. Enforcing public data archiving policies in academic publishing: a study of ecology journals. Big Data Soc. 6 (1) https://doi.org/10.1177/2053951719836258, 205395171983625.

Smith, C., Sokol, E., 2021. ecocomDP: Work with Datasets in the Ecological Community Design Pattern. R Package Version 1.0.0. https://CRAN.R-project.org/package=ecocomDP. (Accessed 10 June 2021).

Stokstad, E., 2011. Open-source ecology takes root across the world. Science 334 (6054), 308. https://doi.org/10.1126/science.334.6054.308.

Tarboton, D.G., Horsburgh, J.S., Maidment, D.R., 2008. CUAHSI Community Observations Data Model (ODM) Version 1.1 Design Specifications.

Thorpe, A.S., Barnett, D.T., Elmendorf, S.C., Hinckley, E.-L.S., Hoekman, D., Jones, K.D., LeVan, K.E., Meier, C.L., Stanish, L.F., Thibault, K.M., 2016. Introduction to the sampling designs of the National Ecological Observatory Network Terrestrial Observation System. Ecosphere 7, e01627. https://doi.org/10.1002/ecs2.1627.

Utz, R.M., Fitzgerald, M.R., Goodman, K.J., Parker, S.M., Powell, H., Roehm, C.L., 2013. The National Ecological Observatory Network: an observatory poised to expand spatiotemporal scales of inquiry in aquatic and fisheries science. Fisheries 38, 26–35. https://doi.org/10.1080/03632415.2013.748551.

Walter, J.A., Shoemaker, L.G., Lany, N.K., Castorani, M.C.N., Fey, S.B., Dudney, J.C., Gherardi, L., Portales-Reyes, C., Rypel, A.L., Cottingham, K.L., Suding, K.N., Reuman, D.C., Hallett, L.M., 2021. The spatial synchrony of species richness and its relationship to ecosystem stability. Ecology. https://doi.org/10.1002/ecy.3486 (accepted).

Welti, E., Joern, A., Ellison, A.M., Lightfoot, D., Record, S., Rodenhouse, N., Stanley, E., Kaspari, M., 2021. Meta-Analyses of Insect Temporal Trends Must Account for the Complex Sampling Histories Inherent to Many Long-Term Monitoring Efforts in revision.

Wickham, H., 2014. Tidy data. J. Stat. Softw. *59* (10), 1–23. https://doi.org/10.18637/jss.v059.i10.

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), e29715. https://doi.org/10.1371/journal.pone.0029715.

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.