NEON USER GUIDE TO PLANT PRESENCE AND PERCENT COVER (DP1.10058.001)

PREPARED BY	ORGANIZATION
Sarah Elmendorf	SCI
Dave Barnett	SCI

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	07/14/2017	Initial Release
B	04/29/2019	Revision to reflect 2018-19 protocol optimization
C	05/26/2020	Included general statement about usage of neonUtilities R package and statement about possible location changes. Updated taxonomy information.
D	10/02/2020	Updated Design Changes section and plot figures.
E	04/28/2022	Added language in section 4 Taxonomy addressing RTE species obfuscation in the data. Updated section 5.3 Data Revision with latest information regarding data release. Updated information regarding the geoNEON package.
F	02/20/2024	Updated text and figures to reflect subplotID naming convention change, added identification history table, and added text promoting the use of the neonPlants package to aggregate multiscale data across tables.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

This document provides an overview of the data included in this NEON Level 1 data product, the quality controlled product generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field, for example, the dry weights of litter functional groups from a single collection event are considered the lowest level (Level 0). Raw data that have been quality checked via the steps detailed herein, as well as simple metrics that emerge from the raw data are considered Level 1 data products.

The text herein provides a discussion of measurement theory and implementation, data product provenance, quality assurance and control methods used, and approximations and/or assumptions made during L1 data creation.

1.2 Scope

This document describes the steps needed to generate the L1 data product Plant presence and percent cover - terrestrial species lists from nested subplots and ocular estimates of percent cover - and associated metadata from input data. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the file, NEON Data Variables for Plant presence and percent cover (DP1.10058.001) (AD[05]), provided in the download package for this data product.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the data collected in the field pertaining to NEON Field and Lab Protocol for Plant Diversity (AD[07]). The raw data that are processed in this document are detailed in the file, NEON Raw Data Validation for Plant presence and percent cover (DP0.10004.001) (AD[04]), provided in the download package for this data product. Please note that raw data products (denoted by 'DP0') may not always have the same numbers (e.g., '10033') as the corresponding L1 data product.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Associated Documents

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.000913	TOS Science Design for Spatial Sampling
AD[03]	NEON.DOC.002652	NEON Level Data Products Catalog
AD[04]	Available with data download	Validation csv
AD[05]	Available with data download	Variables csv
AD[06]	NEON.DOC.000912	TOS Science Design for Plant Diversity
AD[07]	NEON.DOC.014042	NEON Field and Lab Protocol for Plant Diversity
AD[08]	NEON.DOC.000008	NEON Acronym List
AD[09]	NEON.DOC.000243	NEON Glossary of Terms
AD[10]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[11]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[12]	NEON.DOC.001024	NEON Field and Lab Protocol for Canopy Foliage Sampling
AD[13]	Available on NEON data portal	NEON Ingest Conversion Language
AD[14]	Available with data download	Categorical Codes csv

3 DATA PRODUCT DESCRIPTION

The Plant presence and percent cover data product provides vascular species lists and ocular estimates of ground cover for terrestrial plants and ancillary abiotic data from individual sampling bouts. The presence and percent cover of species is documented in square, multi-scale plots. Species and abiotic data are reported at the spatial resolution at which they were observed (see 3.9 Data Relationships for processing), and include information on taxonomy, record-specific uncertainty, nativity, and location. Also included in the expanded package are details of collection of plant voucher materials, frozen tissue samples for archiving, and unknown or morphospecies collection and resolution.

Plant presence and percent cover data may be used to describe species diversity, patterns of plant species invasion, patterns of overlap and similarity within and across NEON sites, and the relationship of particular species or species richness to other ecosystem descriptors as measured by other NEON data products such as vegetation structure, productivity, remote sensing, and ecosystem exchange.

3.1 Spatial Sampling Design

Plant presence and percent cover sampling is conducted at terrestrial NEON sites. Sampling occurs at three Tower Base plots per site, and a variable number of Distributed Base plots depending on the size and heterogeneity of the site. Locations of Tower Base plots are selected within the 90% flux footprint of the primary and secondary airsheds (and additional areas in close proximity to the airshed as necessary to accommodate sufficient spacing between plots). Distributed Base plots are randomly placed within the dominant National Land Cover Database (NLCD) cover types within the site. The number of plots per cover type is proportional to the square root of total area within each land cover type. At some sites, available space, plot spacing requirements, and/or the tower airshed size restricts the number of plots that can be sampled for plant presence and percent cover. Plot edges must be separated by a distance 150% of one edge of the plot (e.g., 40m x 40m Tower Base Plots must be 60m apart); plot centers must be greater than 50m from large paved roads and plot edges must be 10m from two-track dirt roads; plot centers must be 50m from buildings and other non-NEON infrastructure; streams larger than 1m must not intersect plots. See TOS Science Design for Plant Diversity (AD[06]), NEON Field and Lab Protocol for Plant Diversity (AD[07]), TOS Science Design for Spatial Sampling (AD[02]) and for further details.

Within each plot, the presence and percent cover of species and ancillary data is observed in six $1m^2$ subplots (see section 3.3 Design Changes). The presence of species is observed in six $10m^2$ subplots and four $100m^2$ subplots, which can be combined for a list of species at the $400m^2$ plot scale (Figure 1). The multi-scale plot design is consistent with methods of the Carolina Vegetation Project (Peet et al. 1996), similar to other multiscale methods (Stohlgren 2007), and based on Robert Whittaker's approach to sampling vegetation (Smida 1984).

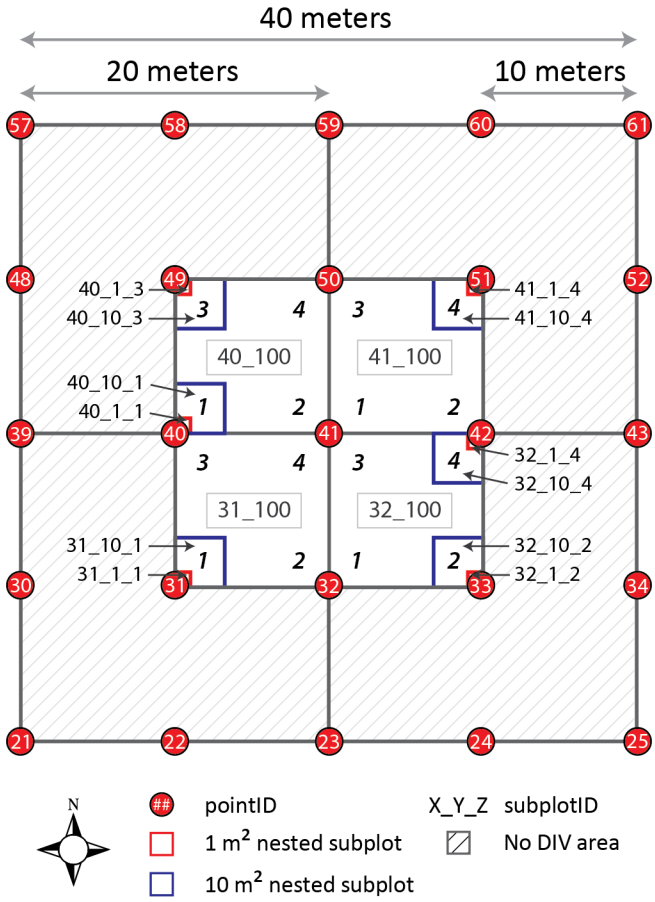

Figure 1: Plot and subplot layout with subplot naming convention and and assoicated points that define the plot permimeter.

As much as possible, sampling occurs in the same locations over the lifetime of the Observatory. However, over time some plot locations may become impossible to sample due to disturbance or other local changes. When this occurs, the location and its location ID are retired. A location may also shift to slightly different coordinates. Refer to the locations endpoint of the NEON API for details about locations that have been moved or retired: https://data.neonscience.org/data-api/endpoints/locations/

3.2 Temporal Sampling Design

Plots are sampled annually at most sites to correspond with the display of diagnostic plant parts such as flowers, fruits, and/or seeds, roughly during peak greenness. The collective observation of each target plot within a site, a single time, and in one year constitutes a 'bout'(field **boutNumber**). A single bout - each plot observed once per year - is completed at most sites, while the few sites with distinct bimodal peaks in greenness and species composition have two bouts - each plot observed twice a year.

The $1m^2$ subplots are observed at every plot every year during each bout. Observations of the larger subplots - six $10m^2$ subplots and four $100m^2$ subplots - are made every other year at each plot within each

site. For example, the entire plot - 1m^2 subplots, 10m^2, and 100m^2 subplots were sampled in 2019 at (STEI), but only the 1m^2 subplots were sampled at all plots at Steigerwaldt the following year in 2020.

3.3 Design Changes

While maintenance of a design that produces comparable data through time is a priority, the observation of plant diversity was subjected to adjustments designed to increase the efficiency and make data more clear to end users. Changes arise due to continual evaluation of the sampling design for best practices and to ensure that allocation of sampling effort is poised to maximize returns to the scientific community. All design changes are made in collaboration with external technical working groups. The following changes have been made to the design:

- 2018: The 10 and 100m^2 observations were eliminated at half of the plots at each site. At each site during each bout (most sites only sample one bout), the complete plot (all 1, 10, and 100m^2 subplots) were observed at half of the plots and only the 1m^2 subplots were observed at the other half of the plots. This was a one-time sampling reduction in response to budget constraints.
- Prior to 2019: The 1m^2 subplots and the 10 and 100m^2 subplots were observed every year at each plot targeted for sampling at each site.
- Prior to 2019: Two additional 1 and 10m^2 subplots were sampled. The presence and percent cover of species and ancillary data was observed in eight 1m^2 subplots. The presence of species was observed in eight 10m^2 subplots and four 100m^2 subplots, which can be combined for a list of species at the 400m^2 plot scale (Figure 2).
- 2020: The field **samplingImpractical** was added to the data to allow for the generation a record when a subplot or entire plot could not be sampled for a particular bout and year. If field sampling was not possible **samplingImpractical** is populated with a value other than 'OK' (e.g., 'location flooded') and no plant species data are recorded.
- 2020: The field **biophysicalCriteria** was added to the data to account for instances when sampling occurred but conditions were not optimal. If conditions were suboptimal - the majority of species present do not possess plant parts conducive to identification - the field **biophysicalCriteria** is populated with a value other than 'OK - no known exceptions' (e.g., 'conditions not met: most plants not yet flowering') but plant species and ancillary data are recorded.
- 2024 Data Release: The subplot naming convention was changed in all data. See section 3.5 Spatial Resolution and Extent for details. Additionally, the table div_identificationHistory was added to the data product to enable tracking of edits to taxonomic determinations.

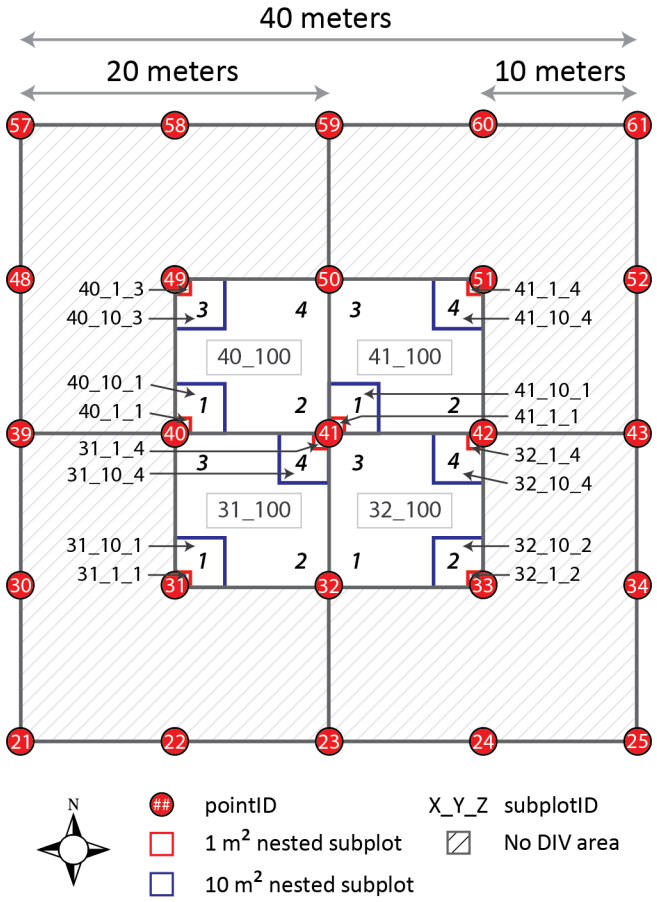

Figure 2: Plot and subplot layout of data gathered prior to 2019; sampling at the 1m^2 and 10m^2 subplots in the center of the plot was discontinued to reduce sampling time and impacts of sampling.

3.4 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Plant presence and percent cover (DP0.10004.001) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the file, NEON Data Variables for Plant presence and percent cover (DP1.10058.001) (AD[05]).

Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 16 February 2014), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 16 February 2014), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), where applicable. NEON TOS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and GEOID09 for its reference gravitational ellipsoid. Latitudes and longitudes are denoted in decimal notation to six decimal places, with longitudes indicated as negative west of the Greenwich meridian.

Some variables described in this document may be for NEON internal use only and will not appear in

downloaded data.

3.5 Spatial Resolution and Extent

The finest resolution at which the Plant Presence and Percent Cover data will be tracked is at the $1m^2$ subplot, resulting in the following spatial hierarchy from finest to coarsest resolution:

- The presence and percent cover of species and ancillary data is observed in six $1m^2$ subplots (e.g., **subplotID** = 40_1_1). The presence of species is observed in six $10m^2$ subplots (e.g., **subplotID** = 40_10_1) and four $100m^2$ subplots (e.g., **subplotID** = 40_100), which can be combined for a list of species at the $400m^2$ plot scale (e.g., **plotID** = CPER_030). Plots are located within NEON sites (e.g., **siteID** = CPER), which are located with NEON domains (e.g., **domainID** = D10).

The naming convention for subplots within Base plots consists of the identity of the plot point in the southwest corner of the subplot, the scale or size of the subplot, and, for those subplots smaller than $100m^2$, the corner of the $100m^2$ subplot in which the smaller subplot is located. For example, '31_100' refers to point 21 in the southwest corner and is $100m^2$ (10 m x 10 m), and '31_1_1' is a subplot with point 31 in the southwest corner, is $1m^2$, and is in corner 1 of the $100m^2$ subplot with point 31 in the southwest corner (Figure 1). Subplots within Base plots in data releases prior to the 2024 data release – Release 2024 – follow a slightly different naming convention. Previously, subplots of $100m^2$ or $400m^2$ were identified by the point in the southwest corner of the subplot (e.g., '21' or '41'). Subplots smaller than $100m^2$ were previously named according to the point in the southwest corner of the $100m^2$ subplot, the subplot corner, and the scale (e.g., '41.2.10'). Differences are:

- The inclusion of scale in all subplotIDs; what was '31' is now '31_100'
- The change in location of the subplot scale or size and the identifying corner in the subplot string for subplots smaller than $100m^2$; what was '31.1.10' is now '31_10_1'.
- String components are now separated by an underscore ("_") instead of a period ("."); what was subplot '31.1.10' will become '31_10_1'.

The basic spatial data included in the data downloaded include the latitude, longitude, and elevation of the centroid of the plot where sampling occurred + associated uncertainty due to GPS error and plot width.

To obtain the location of each subplot center, there are three options:

1. Use the getLocTOS function from the geoNEON package, available here: https://github.com/NEONScience/NEON-geolocation. The **namedLocation** field in the div_1m2Data and div_10m2Data100m2Data tables is the named location of the plot; more precise geographic data require the named location of the subplot (Figure 1). Construct the named location of the subplot of each record in by concatenating the fields for **namedLocation** and **subplotID** as: namedLocation + '.' + subplotID, e.g. subplotID '41_100' of namedLocation 'HARV_052.basePlot.div' has a complete named location of 'HARV_052.basePlot.div.41_100'. Similarly, subplotID '31_1_1' of namedLocation 'MOAB_044.basePlot.div' has a complete named location of 'MOAB_044.basePlot.div.31_1_1'.
2. Use the API (http://data.neonscience.org/api; e.g., http://data.neonscience.org/api/v0/locations/HARV_052.basePlot.div.41_100) to query for elevation (**locationElevation**), easting (**locationUtmEasting**), northing (**locationUtmNorthing**), coordinate

uncertainty (**coordinateUncertainty**), elevation uncertainty (**elevationUncertainty**), utm zone (**locationUtmZone**), latitude (**locationDecimalLatitude**) and longitude (**locationDecimalLatitude**) as inputs.

3. Download spatial files (including shapefiles) from the NEONScience.org document library (https://data.neonscience.org/documents).

Subplot centroids are not obtained in the field but, estimated based on high-resolution GPS retrieval of the spatial coordinates of pointIDs at plots (Figure 1). Due to the challenges of establishing plots in wild vegetation and, in complex topography in many cases, reported **coordinateUncertainty** associated with subplot centroids has been increased:

- 0.25m for $1m^2$ subplot centroids
- 1.0m for $10m^2$ subplot centroids
- 2.0m for $100m^2$ subplot centroids

3.6 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the start and end date (typically 1-3 days) a specific plot was sampled.

3.7 Associated Data Streams

namedLocation and **taxonID** from div_voucher and div_geneticarchive (in the expanded package) are linking variables that tie vouchered species (and individuals if tagID is equal in both) to other data products such as Plant presence and percent cover, Woody plant vegetation structure (DP1.10098.001), Plant phenology observations (DP1.10055.001), and Plant foliar physical and chemical properties (DP1.10026.001).

Species are identified to the lowest taxonomic rank possible (entered in fields **taxonID** and **scientificName** and specified in **taxonRank** in tables div_1m2Data and div_10m2Data100m2Data). In cases where identity can't be resolved in the field but might be in the lab or herbarium, the lowest possible taxonomic rank is recorded and an unknown 'morphospecies' name (in the field **morphospeciesID** in tables div_1m2Data and div_10m2Data100m2Data) is recorded. When resolved, these morphospecies are published in the fields **taxonID** and **scientificName** in table div_morphspecies. Records in the table div_morphospecies can be linked to tables div_1m2Data and div_10m2Data100m2Data by linking on the fields **siteID** and **morphospeciesID**.

The protocol dictates that 10 foliar tissue samples be collected from three different species and requires that the individual from which these tissue samples is collected and vouchered. The samples can be linked with fields in the data of both tables. The table div_voucher has the field **voucherSampleID** (e.g., pla.OAES.20151014.10:30.dtb.V123) and div_geneticarchive contains the field **geneticSampleID** (gen.OAES.20151014.10:30). The combination of the site (e.g., OAES), date (as yyyymmdd, e.g., 20151014) and the time (e.g., 10.30) in both of the **voucherSampleID** and the **geneticSampleID** provide sufficiently unique information by which these samples can be linked. Not all of the genetic tissue sam-

ples have corresponding vouchers, and only a small subset of vouchers have associated (from the same individual or species) tissue samples.

3.8 Product Instances

There are a maximum of two Plant presence and percent cover bouts per year, with data collected from no more than 33 plots per site per bout. Each plot will yield data for no more than eight $1m^2$ subplots (six after 2018), no more than eight $10m^2$ subplots (six after 2018), and four $100m^2$ subplots.

The plant voucher collection will result in approximately 20 specimens per site per year.

The frozen tissues will result in 30 samples per site every five years.

3.9 Data Relationships

When a plot is sampled, there will be at least two records - one record for abiotic variables and one for plant species - for each of the six (eight prior to 2019) $1m^2$ subplots, a minimum of 12 records in div_1m2Data for each unique namedLocation (**siteID** and **plotID**). The presence/absence of either plant species (e.g., **targetTaxaPresent** = 'Y' or 'N') or other abiotic variables (**otherVariablesPresent**) is to be recorded for each $1m^2$ plot surveyed. The table div_10m2Data100m2Data will have one or more records for the six (eight prior to 2019) $10m^2$ subplots, and one or more records for each of the four $100m^2$ subplots. The table div_voucher will only generate records when specimens are collected, and namedLocation may reflect a **plotID** or a **siteID** (in cases where specimens are collected outside a plot). The table div_geneticarchive will generate 30 records per site over three consecutive years. Duplicates and/or missing data may exist where protocol and/or data entry aberrations have occurred; users should check data carefully for anomalies.

Accurate and comprehensive species lists for $10m^2$ and $100m^2$ subplots and entire plots requires data aggregation across scales that accounts for the contribution of species found in nested subplots. For example, a complete list of species present in $10m^2$ subplot 31_10_1 must include the species observed in $1m^2$ subplot 31_1_1 which occur but are not published in $10m^2$ subplot 31_10_1 in the table div_10m2Data100m2Data (Figure 1). Code stored in a GitHub repository (https://github.com/NEONScience/neonPlants) contains the the function stackPlantPresence that should be employed to appropriately aggregate data.

Alternatively, linking presence of species documented in the $1m^2$ subplots and published in table div_1m2Data with species documented in 10 and $100m^2$ subplots published in table div_10m2Data100m2Data requires the correct named location (**siteID** and **plotID**), year, and bout. This sampling event is captured in the field **eventID** (**siteID.boutNumber.year**, e.g., STEI.1.2020) that can be used to link the tables. The protocol dictates that species are observed first in a $1m^2$ subplot (e.g., subplotID 31_1_1 from table div_1m2Data) followed by the recording of instances of new species in the remaining $9m^2$ area of the corresponding $10m^2$ subplot (e.g., subplotID 31_10_1 in table div_10m2Data100m2Data) without requiring the re-recording of species already found in the nested $1m^2$ subplot (e.g., subplotID 31_1_1 in table div_1m2Data). Because the list of species published in the $10m^2$ nested subplot does not include species found in the nested $1m^2$ subplot, the species reported at the $10m^2$ subplot (e.g., subplotID 31_10_1 in table div_10m2Data100m2Data) must be combined with the

species reported at the nested 1m^2 subplot (e.g., subplotID 31_1_1 in table div_1m2Data) to create the complete list of species observed in the 10m^2 subplot (e.g., subplotID 31_10_1). Similarly, generating lists of species at a 100m^2 subplot (e.g., subplotID 31_100 from in div_10m2Data100m2Data) must be generated by appropriately combining the species published in the one or two nested 1m^2 subplots (e.g., subplotID 32_1_2 and 32_1_4 in table div_1m2Data), the one or two nested 10m^2 subplots (e.g., subplotID 32_10_2 and 32_10_4 from table div_10m2Data100m2Data), and 100m^2 subplot (e.g., subplotID 32_100 from table div_10m2Data100m2Data). Table 1 provides the logic for all nested subplots. In cases where the sampling in nested subplots captured all of the species in the larger subplot (e.g., all species found in subplot 41_100 were contained in subplots 41_1_4 and 41_10_4, data for subplot 41_100 will be recorded as targetTaxaPresent = 'Y' – to denote the presence of plants in the 100m^2 subplot – and additionalSpecies = 'N'– to denote that no new species were observed in the 90m^2 not occupied by the nested subplots. Generating a species list for the entire 400m^2 plot requires the combination of all subplots where the **plotID** and **eventID** are equal.

Species that were not identified in the field but possessing plant parts thought to be conducive to identification with additional resources (e.g., herbaria, taxonomic experts) are captured in the table div_morphospecies. A unique morphospeciesID can be documented in multiple subplots and plots. Data in div_1m2Data and div_10m2Data100m2Data can be updated by linking on the **siteID** and **morphospeciesID**.

The table div_identificationHistory is designed to track to updates to taxanomic determinations (see section 4, Taxonomy). Records are only created when data corrections to taxonomic identifications are made. If errors in identification are detected through QAQC processes, corrected taxonomy will be provided in the table to which the change is relevant. The table div_identificationHistory is populated with all prior names used for species in the data product. When data are populated in the div_identificationHistory table, **identificationHistoryID** is used as a linking variable between the div_identificationHistory table and all other tables in this data product where updates were made.

Table 1: Nested subplots necessary to collate full species lists. It is advisable to aggregate data with the function stackPlantPresence in the package neonPlants (https://github.com/NEONScience/neonPlants).

subplotID	**full list contained in:**
31_10_1	31_10_1 and 31_1_1
31_10_4	31_10_4 and 31_1_4
31_100	31_10_1 and 31_1_1 and 31_10_4 and 34_1_4
32_10_2	32_10_2 and 32_1_2
32_10_4	32_10_4 and 32_1_4
32_100	32_10_2 and 32_1_2 and 32_10_4 and 32_1_4
40_10_1	40_10_1 and 40_1_1
40_10_3	40_10_3 and 40_1_3
40_100	40_10_1 and 40_1_1 and 40_10_3 and 40_1_3
41_10_1	41_10_1 and 41_1_1
41_10_4	41_10_4 and 41_1_4
40_100	41_10_1 and 41_1_1 and 41_10_4 and 41_1_4
entire 400m^2 plot	31_10_1 and 31_1_1 and 31_10_4 and 31_1_4 and 32_10_2 and 32_1_2 and 32_10_4 and 32_1_4 and 40_10_1 and 40_1_1 and 40_10_3 and 40_1_3 and 41_10_1 and 41_1_1 and 41_10_4 and 41_1_4

div_1m2Data -> One record expected per taxonID present in a given subplot per plotID per bout plus one record expected per class of ground cover present in a given **subplotID** per **plotID** per **boutID** (e.g. litter, lichen).

- primary key -> plotID, subplotID, boutNumber, eventID, taxonID, identificationQualifier, morphospeciesID, otherVariables

div_10m2Data100m2Data -> One record expected per taxonID present in a given subplot per plotID per bout for each taxon that is NOT already recorded in a nested subplot

- primary key -> plotID, subplotID, boutNumber, eventID, taxonID, identificationQualifier, morphospeciesID

div_morphospecies -> one record for each species identified at a later date

- primary key -> siteID, morphospeciesID

div_voucher -> Records only generated when specimen collected

- primary key -> voucherSampleID, voucherSampleCode

div_geneticarchive -> Thirty records every five years per site

- primary key -> geneticSampleID, geneticSampleCode

div_identificationHIstory -> one record for each change to taxonomy

- primary key -> identificationHistoryID, identifiedDate

Data are most easily acquired and stacked with the loadByProduct function in the neonUtilities R package (CRAN; https://cran.r-project.org/web/packages/neonUtilities/index.html) and then aggregated with the stackPlantPresence function in the neonPlants package (https://github.com/NEONScience/neonPlants). Data downloaded from the NEON Data Portal are provided in separate data files for each site and month requested. The neonUtilities R package contains functions to merge these files across sites and months into a single file for each table described above. The neonUtilities package is available from the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/web/packages/neonUtilities/index.html) and can be installed using the install.packages() function in R. For instructions on using neonUtilities to merge NEON data files, see the Download and Explore NEON Data tutorial on the NEON website: https://www.neonscience.org/download-explore-neon-data

4 TAXONOMY

NEON manages taxonomic entries by maintaining a master taxonomy list based on the community standard, if one exists. Through the master taxonomy list, synonyms submitted in the data are converted to the appropriate name in use by the standard. The master taxonomy for plants is the USDA PLANTS Database (USDA, NRCS. 2014. https://plants.usda.gov). Taxon ID codes used to identify taxonomic concepts in the NEON master taxonomy list are alpha-numeric codes, 4-6 characters in length based on the accepted scientific name. Each code is composed of the first two letters of the genus, followed by the first two letters of the species and first letter of the terminal infraspecific name (if applicable) then, if needed, a tiebreaking number to address duplicate codes. Genus and family symbols are the first five (genus) or six (family) letters of the name, plus tiebreaking number (if needed). Symbols were first used in the Soil Conservation Service's National List of Scientific Plant Names (NLSPN) and have been perpetuated in the PLANTS system. The portions of the PLANTS Database included in the NEON plant master taxonomy list includes native and naturalized plants present in NEON observatory sampling area including the Lower 48 U.S. States, Alaska, Hawaii, and Puerto Rico. NEON plans to keep the taxonomy updated in accordance with USDA PLANTS Database starting in 2020 and annually thereafter.

The master taxonomy list includes geographic range and nativity as described by the USDA PLANTS Database. A list for each NEON domain includes those species with ranges that overlap the domain as well as nativity designations - introduced or native - in that part of the range. Errors are generated if a species is reported at a location outside of its known range. If the record proves to be a reliable report, the master taxonomy table is updated to reflect the distribution change.

Prior to the 2022 data release, publication of species identifications were obfuscated to a higher taxonomic rank when the taxon was found to be listed as threatened, endangered, or sensitive at the state level where the observation was recorded. The state-level obfuscation routine was removed from the data publication process at all locations excluding sites located in D01 and D20, and data have been reprocessed to remove the obfuscation of state-listed taxa for all years. Federally listed threatened and endangered or sensitive species remain obfuscated at all sites and sensitive species remain redacted at National Park sites.

The full master taxonomy lists are available on the NEON Data Portal for browsing and download: http://data.neonscience.org/static/taxon.html.

4.1 Identification History

Beginning in 2022, the identificationHistory table was added to track any changes to taxonomic identifications that have been published in NEON data. Such taxonomic revisions may be necessary when errors are found in QAQC checks or when evidence from genetic analysis of samples or re-analysis of archived samples indicate a revision is necessary. Requests for taxonomic changes are reviewed by NEON science staff. Proposed changes are evaluated based on evidence in the form of photographs, existing samples, genetic data, consultation with taxonomic experts, or range maps. Upon approval, the existing record in all tables associated with this data product is updated with the new taxonomic information and a unique identifier is added to the identificationHistoryID field. A record with the same **identificationHistoryID** is created in the div_identificationHistory where the previous taxonomic information is archived along with the date the change was made.

5 DATA QUALITY

5.1 Data Entry Constraint and Validation

Many quality control measures are implemented at the point of data entry within a mobile data entry application or web user interface (UI). For example, data formats are constrained and data values controlled through the provision of dropdown options which reduces the number of processing steps necessary to prepare the raw data for publication. An additional set of constraints are implemented during the process of ingest into the NEON database. The product-specific data constraint and validation requirements built into data entry applications and database ingest are described in the document NEON Raw Data Validation for Plant presence and percent cover (DP0.10004.001), provided with every download of this data product. Contained within this file is a field named **entryValidationRulesForm**, which describes syntactically the validation rules for each field built into the data entry application. Data entry constraints are described in Nicl syntax in the validation file provided with every data download, and the Nicl language is described in NEON's Ingest Conversion Language (NICL) specifications ([AD[11]).

A schematic of the data entry application design is depicted in Figure 3.

5.2 Automated Data Processing Steps

Following data entry into a mobile application of web user interface, the steps used to process the data through to publication on the NEON Data Portal are detailed in the NEON Algorithm Theoretical Basis Document: OS Generic Transitions (AD[11]).

5.3 Data Revision

All data are provisional until a numbered version is released. Annually, NEON releases a static version of all or almost all data products, annotated with digital object identifiers (DOIs). The first data Release was made in 2021. During the provisional period, QA/QC is an active process, as opposed to a discrete

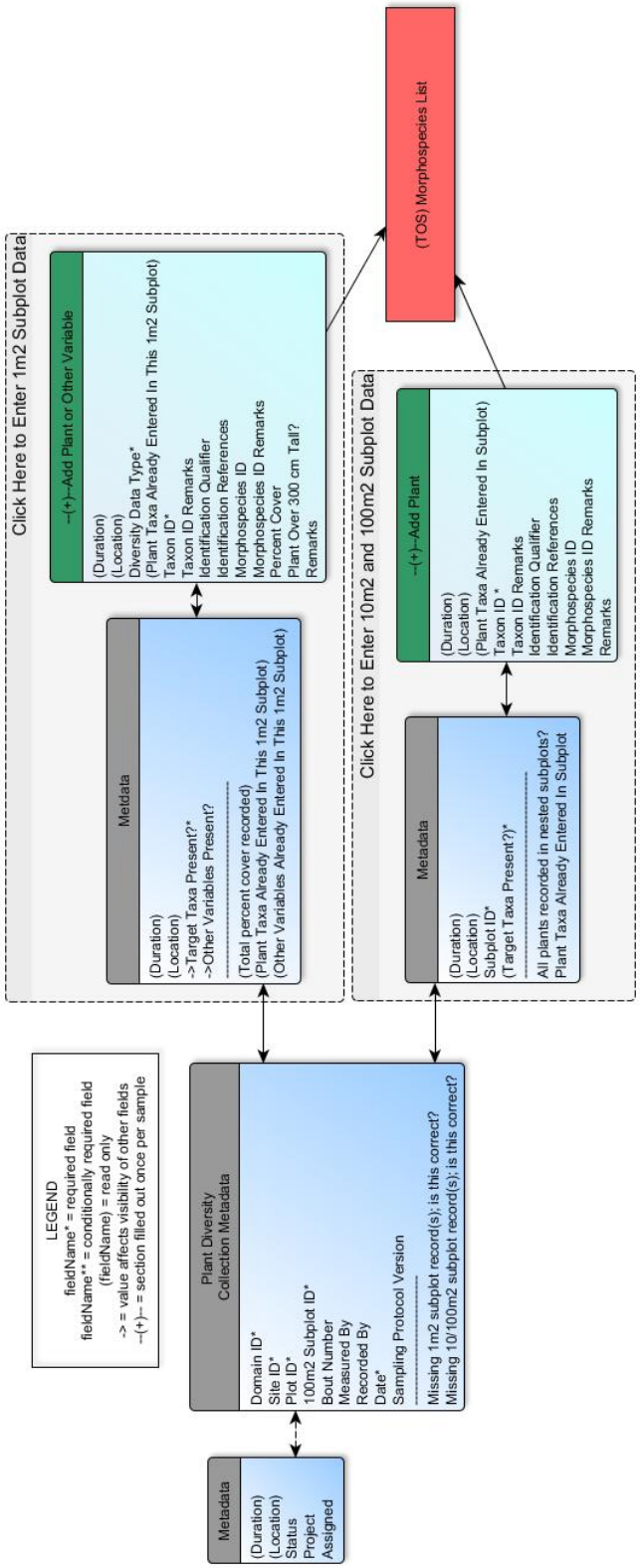

Figure 3: Schematic of the applications used by field technicians to enter plant presence and percent cover data

activity performed once, and records are updated on a rolling basis as a result of scheduled tests or feedback from data users. The Issue Log section of the data product landing page contains a history of major known errors and revisions.

5.4 Quality Flagging

The **dataQF** field in each data record is a quality flag for known errors applying to the record. There are currently no dataQF codes in use in this data product.

Prior to 2017, data was collected with a system in the field that did not include the full suite of front-end quality assurance checks applied to data for the 2017 and subsequent sampling years.

Records of land management activities, disturbances, and other incidents of ecological note that may have a potential impact are found in the Site Management and Event Reporting data product (DP1.10111.001)

6 REFERENCES

Peet, R. K., T. R. Wentworth, and P. S. White. 1998. A flexible, multipurpose method for recording vegetation composition and structure. Castanea 63(3):262-274.

Shmida, A. 1984. Whittaker's plant diversity sampling method. Israel Journal of Botany 33:41-46.

Stohlgren, T. J. 2007. Measuring plant diversity, lessons from the field. Oxford University Press, New York.

USDA, NRCS. 2014. The PLANTS Database (http://plants.usda.gov, 25 August 2014). National Plant Data Team, Greensboro, NC 27401-4901 USA.