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

PREPARED BY	ORGANIZATION
Sarah Elmendorf	DPS
Dave Barnett	TOS

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.

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 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 material; frozen tissue samples for archiving; and morphospecies collection and resolution across all terrestrial plant protocols.

Plant presence and percent cover data may be used to describe patterns of invasion, species diversity, 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, and ecosystem exchange.

3.1 Spatial Sampling Design

Plant presence and percent cover sampling is conducted at terrestrial NEON sites. Sampling is conducted at three tower plots per site, plus a variable number of distributed base plots, depending on the size and heterogeneity of the site. Locations of tower 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 selected within the dominant National Land Cover Database (NLCD) cover types within the site, with the number of plots per cover type allocated proportional to the square root of total area within each land cover type. In some cases, 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. Specifically, 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.

The presence and percent cover of species and ancillary data is observed in six $1m^2$ subplots. 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).

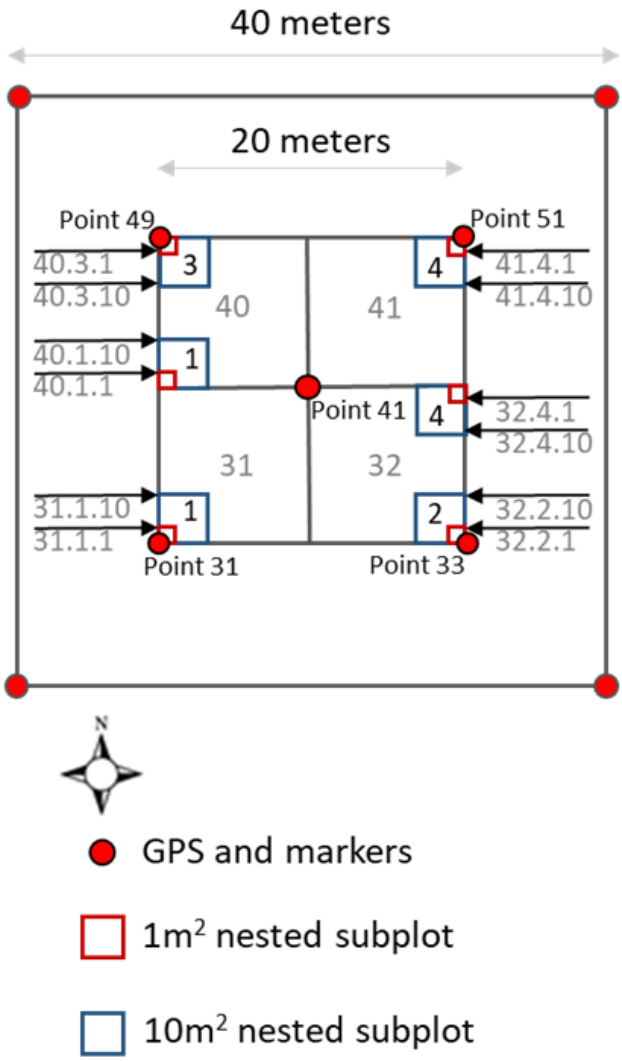

Figure 1: Plot and subplot layout describing subplot numbering and points (31, 33, 49, 51) where GPS locations are recorded.

As much as possible, sampling occurs in the same locations over the lifetime of the Observatory. However, over time some sampling plots may become impossible to sample, due to disturbance or other local changes. When this occurs, the location and its location ID are retired. A new location is established, and given a new location ID. Refer to the data product change log for details about plots that have been retired.

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. Select sites with bimodal peaks in greenness and species composition are sampled twice a year.

The $1m^2$ subplots at which the presence and percent cover of species and ancillary data is 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, including subplot were sampled in 2019 at Steigerwaldt (STEI), but only the $1m^2$ subplots were sampled at all plots at the site the following year in 2020.

3.3 Design Changes

While the sampling approach has remained consistent to produce comparable data, the observation of plant diversity was subjected to changes to increase the efficiency of sampling and make data more clear to end users. There have been several design changes that have been implemented over the course of data collection. Such changes arise due to continual evaluation of the sampling design for best practices in collaboration with external technical working groups. Design changes also occur when optimization of the is necessary to ensure that allocation of sampling effort is poised to maximize returns to the scientific community.

- 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.
- 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.

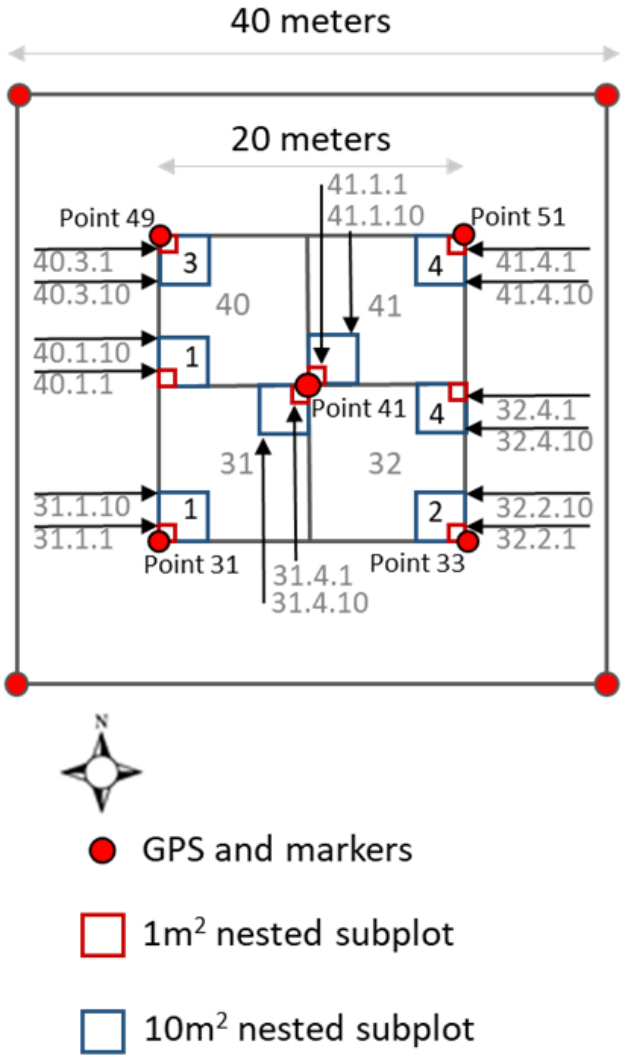

Figure 2: Plot and subplot layout describing design changes.

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 hierarch 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), 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 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. Shapefiles of all NEON Terrestrial Observation System sampling locations can be found in the Document Library: http://data.neonscience.org/documents.

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

1. Use the def.calc.geo.os function from the geoNEON package, available here: https://github.com/NEONScience/NEON-geolocation
2. The namedLocation field in the div_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' of namedLocation 'HARV_052.basePlot.div' has a complete named location of 'HARV_052.basePlot.div.41'.

3. Use the API (http://data.neonscience.org/api; e.g. http://data.neonscience.org/api/v0/locations/HARV_052.basePlot.div.41) to query for elevation("locationElevation"), easting("locationUtmEasting"), northing("locationUtmNorthing"), coordinateUncertainty ("Value for Coordinate uncertainty"), elevationUncertainty ("Value for Elevation uncertainty"), and utmZone ("locationUtmZone"), latitude ("locationDecimalLatitude") and longitude ("locationDecimalLatitude") as inputs to the next step.
4. Increase coordinateUncertainty by an appropriate amount to account for error introduced by navigating within plots. Additional error may be introduced due to tape stretching to navigate to locations within plots and is estimated as:

- 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_voucherand 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).

The protocol dictates that species which could not be identified in the field 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) and recorded with an unknown or 'morphospecies' name (in morphospeciesID in tables div_1m2Data and div_10m2Data100m2Data). These morphospeciesID records will be published periodically, with resolved taxonomic determinations when possible. When morphospecies are resolved, a combination of the linking variables morphospeciesID and measuredBy allow unknown species designated as morphospecies in the field data (div_1m2Data and div_10m2Data100m2Data) to be updated with resolved identities.

The protocol dictates that 10 foliar tissue samples be collected from three different species and requires that the individual from which one of these tissue samples is collected is also vouchered. div_voucher_pub has the field voucherSampleID (e.g., pla.OAES.20151014.10:30.dtb.V123) and div_geneticarchive_pub 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.

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, the protocol dictates that 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_pub for each unique namedLocation (site and plotID). The presence/absence of either plant species (e.g., targetTaxaPresent = 'Y' or 'N') or other abiotic variables (otherVariablesPresent) is 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 only during intermittent sampling years.

Generating a comprehensive list of plant species for an entire plot requires the aggregatoion of tables. Appropriately 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_pub requires the correct named location (siteID and plotID), year, and bout. This sampling event is captured in the field **eventID** (plotID.boutNumber.year, e.g., STEI_001.1.2020) that can be used to link the tables. If this field is not populated, it can be created by aggregating the fields **plotID**, **boutNumber**, and year (subset from **endDate**).

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 remaining $9m^2$ area of the corresponding $10m^2$ subplot (e.g., subplotID 31.1.10 from table div_10m2Data100m2Data_pub) without requiring the re-recording of species already found in the nested $1m^2$ area. The list of species published in the $10m^2$ nested subplot reflects species not found in the nested $1m^2$ subplot; the species reported at the $10m^2$ subplot (e.g., subplotID 31.1.10 from table div_10m2Data100m2Data) must be combined with the species reported at the nested $1m^2$ subplot (e.g., subplotID 31.1.1 from table div_1m2Data_pub) to create the complete list of species observed in the $10m^2$ subplot (e.g., subplotID 31.1.10). Similarly, generating lists of species at a $100m^2$ subplot (e.g., subplotID 31 from table table div_10m2Data100m2Data) must be generated by appropriately combining the species published in the two nested $1m^2$ subplots (e.g., subplotID 31.1.1 and 31.4.1 from table div_1m2Data), the two nested $10m^2$ subplots (e.g., subplotID 31.1.10 and 31.4.10 from table div_10m2Data100m2Data), and $100m^2$ subplot (e.g., subplotID 31 from table table div_10m2Data100m2Data). Table 1 provides the logic for all nested subplots. In cases where the sampling in the corresponding nested subplots captured all of the species in the larger sub-

plot (e.g. all species found in subplot 31 were contained in subplots 31.1.10, 31.4.10, 31.1.1, 31.4.1, data for subplot 31 will be recorded as targetTaxaPresent = 'Y' – to denote the presence of plants – and additionalSpecies = 'N'– to denote that no further additions to the taxon list were necessary at this scale). Generating a species list for the entire 400m^2 plot requires the combination of all subplots where the namedLocation (plotID) are equal.

Table 1: Nested subplots necessary to collate full species lists

subplotID	**full list contained in:**
31.1.10	31.1.10 + 31.1.1
31.4.10	34.4.10 + 34.4.1
31	31.1.10 + 31.1.1 + 34.4.10 + 34.4.1
32.2.10	32.2.10 + 32.2.1
32.4.10	32.4.10 + 32.4.1
32	32.2.10 + 32.2.1 + 32.4.10 + 32.4.1
40.1.10	40.1.10 + 40.1.1
40.3.10	40.3.10 + 40.3.1
40	40.1.10 + 40.1.1 + 40.3.10 + 40.3.1
41.1.10	41.1.10 + 41.1.1
41.4.10	41.4.10 + 41.4.1
40	41.1.10 + 41.1.1 + 41.4.10 + 41.4.1
entire 400m^2 plot	31.1.10 + 31.1.1 + 34.4.10 + 34.4.1 + 32.2.10 + 32.2.1 + 32.4.10 + 32.4.1 + 40.1.10 + 40.1.1 + 40.3.10 + 40.3.1 + 41.1.10 + 41.1.1 + 41.4.10 + 41.4.1

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 bout (e.g. litter, lichen).

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

div_voucher -> Records only generated when specimen collected.

div_geneticarchive -> Thirty records every five years per site.

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.

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

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; the first release of a static version of NEON data, annotated with a globally unique identifier, is planned to take place in 2020. During the provisional period, QA/QC is an active process, as opposed to a discrete activity performed once, and records are updated on a rolling basis as a result of scheduled tests or feedback from data users. The Change Log section of the data product readme, provided with every data download, 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.

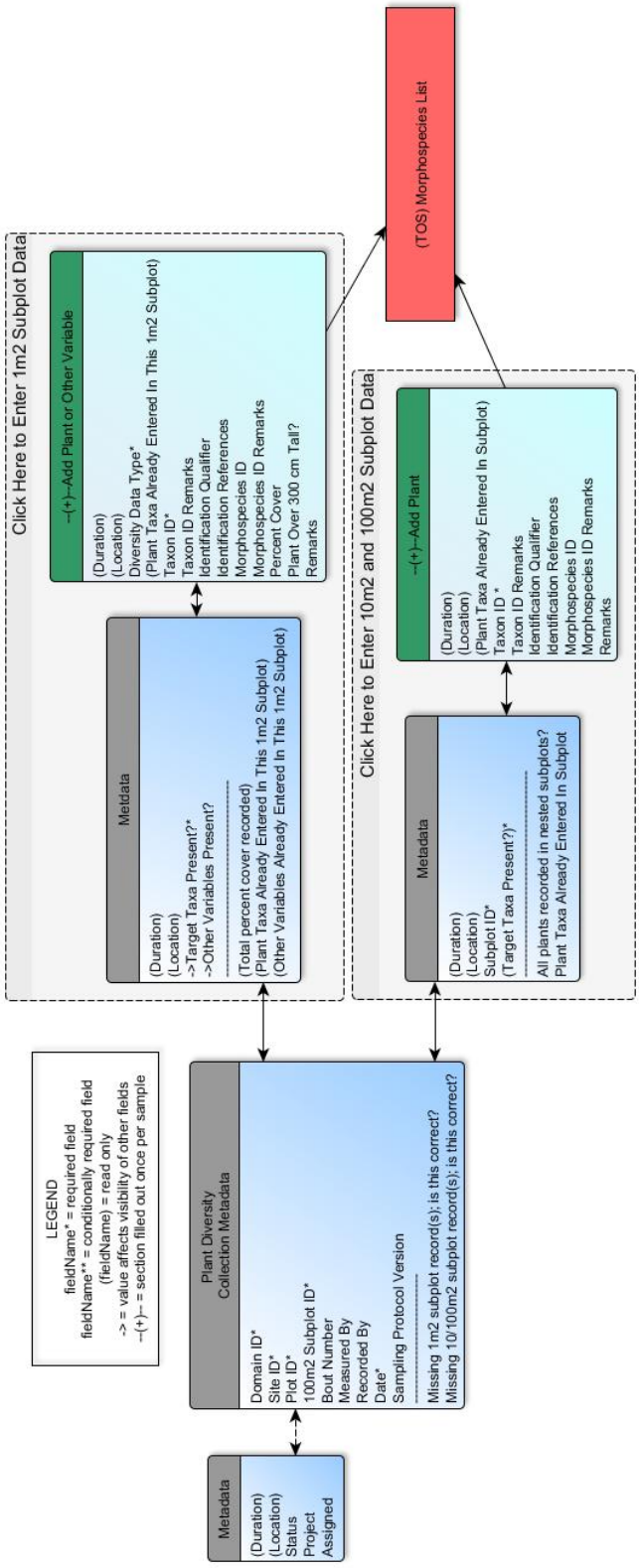

Figure 3: Schematic of the applications used by field technicians to enter plant presence and percent cover data