NEON USER GUIDE TO DNA BARCODING DATA PRODUCTS (NEON.DP1.10020, NEON.DP1.10038, NEON.DP1.10076, NEON.DP1.20105)

PREPARED BY	ORGANIZATION	DATE
Katherine LeVan	SCI	11/28/2017

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	11/22/2017	Initial Release

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 trapping records of individuals 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 four DNA sequence L1 data products:

- Ground beetle sequences DNA barcode data product
- Mosquito sequences DNA barcode data product
- Small mammal sequences DNA barcode data product
- Fish sequences DNA barcode data product

For most data products, NEON provides both data and metadata directly on the NEON data portal. However, because of large community support and development of sequence archives (e.g., Barcode of Life Database, Sequence Read Archive), NEON is externally hosting DNA sequence data for these four products on the Barcode of Life Database (BOLD). Downloads of these products from the NEON data portal will only return sequence metadata, which are also crossposted on BOLD with the sequence data. The format of these data products is designed to output metadata required to upload DNA sequence data (cytochrome oxidase I) on the Barcode of Life Database; to download sequence information, search BOLD with the institution set to 'National Ecological Observatory Network, United States'.

This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the files provided in the download package for these data products (NEON Data Variables for Ground beetle sequences DNA barcode (NEON.DP1.10020) (AD[08]), NEON Data Variables for Mosquito sequences DNA barcode (NEON.DP1.10038) (AD[09]), NEON Data Variables for Small mammal sequences DNA barcode (NEON.DP1.10076.001) (AD[10]), NEON Data Variables for Fish sequences DNA barcode (NEON.DP1.20105.001) (AD[11])). Processed data for the Ground beetle sequences DNA barcode product are generated from TOS Protocol and Procedure: Ground Beetle Sampling (AD[17]), data for the Mosquito sequences DNA barcode product are generated from TOS Protocol and Procedure: Mosquito Sampling (AD[18]), and data for the Small mammal sequences DNA barcode product are generated from TOS Protocol and Procedure: Small Mammal Sampling (AD[19]). Processed data for the Fish sequences DNA barcode product are generated from two protocols: AOS Protocol and Procedure: Fish Sampling in Wadeable Streams (AD[20]) and AOS Protocol and Procedure: Fish Sampling In Lakes (AD[21]). The raw data that are processed in this document are detailed in each ingest file provided in the download package for this data product. The NEON Raw Data Validation for Ground

beetles sampled from pitfall traps (NEON.DP0.10022) (AD[04]) pertains to Ground beetle sequences DNA barcode, NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[05]) pertains to Mosquito sequences DNA barcode, NEON Raw Data Validation for TOS Small Mammal Abundance and Diversity (NEON.DP0.10001) (AD[06]) pertains to Small mammal sequences DNA barcode, and NEON Raw Data Validation for Fish electrofishing, gill netting, and/or fyke netting counts (NEON.DP0.20107) (AD[07]) pertains to Fish sequences DNA barcode. Please note that raw data products (denoted by 'DP0') may not always have the same numbers (e.g., '10003') 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 1, Level 2, and Level 3 Data Products Catalog
AD[04]	NEON.DP0.10003.001 _dataValidation.csv	NEON Raw Data Validation for Ground beetles sampled from pitfall traps (NEON.DP0.10022)
AD[05]	NEON.DP0.10043.001 _dataValidation.csv	NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043)
AD[06]	NEON.DP0.10001.001 _dataValidation.csv	NEON Raw Data Validation for TOS Small Mammal Abundance and Diversity (NEON.DP0.10001)
AD[07]	NEON.DP0.20107.001 _dataValidation.csv	NEON Raw Data Validation for Fish electrofishing, gill netting, and/or fyke netting counts (NEON.DP0.20107)
AD[08]	NEON.DP1.10020.001 _variables.csv	NEON Data Variables for Ground beetle sequences DNA barcode (NEON.DP1.10020)
AD[09]	NEON.DP1.10038.001 _variables.csv	NEON Data Variables for Mosquito sequences DNA barcode (NEON.DP1.10038)
AD[10]	NEON.DP1.10076.001 _variables.csv	NEON Data Variables for Small mammal sequences DNA barcode (NEON.DP1.10076.001)
AD[11]	NEON.DP1.20105.001 _variables.csv	NEON Data Variables for Fish sequences DNA barcode (NEON.DP1.20105.001)
AD[12]	NEON.DOC.000909	TOS Science Design for Ground Beetle Abundance and Diversity
AD[13]	NEON.DOC.000910	TOS Science Design for Mosquito Abundance, Diversity and Phenology
AD[14]	NEON.DOC.000911	TOS Science Design for Vectors and Pathogens
AD[15]	NEON.DOC.000915	TOS Science Design for Small Mammal Abundance and Diversity
AD[16]	NEON.DOC.001152	NEON Aquatic Sample Strategy Document
AD[17]	NEON.DOC.014050	TOS Protocol and Procedure: Ground Beetle Sampling
AD[18]	NEON.DOC.014049	TOS Protocol and Procedure: Mosquito Sampling
AD[19]	NEON.DOC.000481	TOS Protocol and Procedure: Small Mammal Sampling
AD[20]	NEON.DOC.001295	AOS Protocol and Procedure: Fish Sampling in Wadeable Streams
AD[21]	NEON.DOC.001296	AOS Protocol and Procedure: Fish Sampling In Lakes
AD[22]	NEON.DOC.000008	NEON Acronym List
AD[23]	NEON.DOC.000243	NEON Glossary of Terms

AD[24]	OS_Generic_Transitions.pdf	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[25]	Nicl Language.pdf	NEON's Ingest Conversion Language (NICL) specifications

3 DATA PRODUCT DESCRIPTION

The DNA barcoding data products provide sequence data and taxonomic identifications for a subset of ground beetle, mosquito, small mammal and fish individuals that are captured as part of NEON's regular sampling protocols. The DNA barcoding procedure involves the removal tissue, extracting and sequencing DNA from the tissue, and matching that sequence data to sequences from previously identified voucher specimens. Most taxonomic identifications conducted by NEON will be solely determined based on morphological rather than genetic data. Training materials and a comprehensive voucher collection facilitate accurate morphological identifications in the majority of cases. However, identifications of species that are rare, cryptic or for which keys have not been developed, will necessarily be of lower quality. NEON will mitigate potential errors in parataxonomist and taxonomist identifications by sending tissue from a subset of individuals annually to external facilities for DNA sequencing of the Folmer region of the COI gene (aka DNA barcode; Folmer et al. 1994, Hebert et al. 2003). This 'DNA barcode' region has been shown to be effective for use in the identification of most animal taxa, with few exceptions.

Identifications provided by sequence data will improve the quality of technician-derived classifications in the future. Following DNA sequencing, any vouchers associated with the barcoded individuals will be returned to the domain lab of origin and will enhance the voucher collections used by the technicians when making their initial taxonomic assessment. This positive feedback loop will allow technicians to compare newly acquired specimens to a growing collection of high quality vouchers, thereby ensuring increasing accuracy in the identification of new specimens through time. Furthermore, sequencing will also improve the ability of the broader scientific community to make accurate identifications. As NEON accumulates and publishes sequence data on specimens that have also been identified by experts, the quality and quantity of sequence information available for many species will grow. Publicly-available DNA reference sequences will aid in understanding the inter- and intra-specific variation within populations, support accurate identification of specimens by non-experts and reveal the presence of cryptic species.

3.0.1 Ground beetle sequences DNA barcode

Beetles that are rare, particularly difficult to identify, or poorly represented in previous collection events are prioritized for DNA sequencing. DNA sequence data from the Folmer region COI will supplement expert identifications, in order to support the correct taxonomic classification of difficult-to-identify species. See the NEON User Guide to Ground beetles sampled from pitfall traps (NEON.DP1.10022) for additional details about beetle collection and data relationships. For additional details on the sampling design and associated protocol, see also the TOS Science Design for Ground Beetle Abundance and Diversity (AD[12]) and TOS Protocol and Procedure: Ground Beetle Sampling (AD[17]).

Only beetle specimens that have been identified by an expert taxonomist are eligible for DNA barcoding. A subset of specimens that receive DNA barcoding may also be photographed. Data from all barcoded mosquitoes (se-

quence data, location metadata, and photos) are available on the Barcode of Life Database for public use. The **individualID** of all barcoded specimens allows the end-user to connect NEON ground beetle data with the barcoding data **sampleid** supplied via the Barcode of Life Database.

3.0.2 Mosquito sequences DNA barcode

Mosquitoes that are rare, particularly difficult to identify, or poorly represented in previous collection events are prioritized for DNA sequencing. DNA sequence data from the Folmer region COI will supplement expert identifications, in order to support the correct taxonomic classification of difficult-to-identify species. See the NEON User Guide to Mosquitoes sampled from CO2 traps (NEON.DP1.10043) and Mosquito-borne pathogen status (NEON.DP1.10041) for additional details about mosquito collection and data relationships. For additional details on the sampling design and associated protocol, see the TOS Science Design for Mosquito Abundance, Diversity and Phenology (AD[13]) and TOS Protocol and Procedure: Mosquito Sampling (AD[18]).

Only mosquito specimens that have been identified by an expert taxonomist will be eligible for DNA barcoding. A subset of specimens that receive DNA barcoding may also be photographed. Data from all barcoded mosquitoes (sequence data, location metadata, and photos) are available on the Barcode of Life Database for public use. The **individualID** of all barcoded specimens will allow the end-user to connect NEON mosquito data with the barcoding data **sampleid** supplied via the Barcode of Life Database.

3.0.3 Small mammal sequences DNA barcode

Small mammals that are more difficult to identify will be prioritized for DNA sequencing as a Quality Assurance check on technician identifications in the field. DNA sequence data from the Folmer region COI will supplement technician identifications, in order to support the correct taxonomic classification of difficult-to-identify species. See NEON User Guide to Small Mammal Box Trapping (NEON.DP1.10072) for additional details about collection methods. For additional details on the sampling design and associated protocol, see the TOS Science Design for Small Mammal Abundance and Diversity (AD[15]) and TOS Protocol and Procedure: Small Mammal Sampling (AD[19]).

Data from all barcoded small mammals (sequence data and location metadata) are available on the Barcode of Life Database for public use. The **earSampleID**s of all barcoded specimens will allow the end-user to connect NEON mammal data with the barcoding data **sampleid** supplied via the Barcode of Life Database.

3.0.4 Fish sequences DNA barcode

Fish that are more difficult to identify are prioritized for DNA sequencing. DNA sequence data from the Folmer region COI will supplement technician identifications, in order to support the correct taxonomic classification of difficult-to-identify species. See NEON User Guide to NEON User Guide to Fish electrofishing, gill netting, and/or fyke netting counts (NEON.DP1.20107) for additional details about collection methods. For additional details on the sampling design and associated protocol, see the NEON Aquatic Sample Strategy Document (AD[16]), AOS Protocol and Procedure: Fish Sampling in Wadeable Streams (AD[20]) and AOS Protocol and Procedure: Fish Sampling In Lakes (AD[21]).

Data from all barcoded fish (sequence data and location metadata) are available on the Barcode of Life Database for public use. The **dnaSampleID**s of all barcoded specimens will allow the end-user to connect NEON fish data with the barcoding data **sampleid** supplied via the Barcode of Life Database.

3.1 Spatial Sampling Design

Briefly, beetle, mosquito and small mammal sampling is executed at all terrestrial NEON sites and follows a spatially-balanced stratified random design (AD[02]). Beetles are sampled at replicate traps at 10 distributed plots per site. Mosquitoes are sampled at 10 mosquito points per site. Mammal box traps are arrayed in three to eight (depending on the size of the site) 10 x 10 grids, and are collocated with Distributed Base Plots (at which plant, ground beetle and soil sampling may occur), where possible. Fish sampling occurs at all NEON Aquatic sites and captures data on fish within wadeable streams and lakes. Individuals are selected for DNA barcoding without consideration of the particular location within a site from which they were collected.

For additional details on the sampling designs, see also the TOS Science Design for Ground Beetle Abundance and Diversity (AD[12]), TOS Science Design for Mosquito Abundance, Diversity and Phenology (AD[13]),TOS Science Design for Vectors and Pathogens (AD[14]),TOS Science Design for Small Mammal Abundance and Diversity (AD[15]), and NEON Aquatic Sample Strategy Document (AD[16]).

3.2 Temporal Sampling Design

Briefly, beetle, mosquito, small mammals and fish are sampled during timeframes when individuals are active. Beetles are sampled continuously in two-week intervals throughout the growing season at each site up to 13 bouts annually (e.g., up to 6 months of pitfall trapping). Mosquitoes are sampled every two weeks at Core sites and every four weeks at Relocatable sites whenever mosquitoes are actively flying (i.e., when temperatures are sufficiently elevated); this collection may result in up to 26 bouts of sampling annually. Mammals are sampled during 4 bouts annually at Relocatable sites and 6 bouts annually at Core sites. Fish sampling occurs two times per year (once in Spring and once in Fall). Individuals are selected for DNA barcoding without consideration of the particular time period from which they were collected.

For additional details on the sampling designs, see also the TOS Science Design for Ground Beetle Abundance and Diversity (AD[12]), TOS Science Design for Mosquito Abundance, Diversity and Phenology (AD[13]),TOS Science Design for Vectors and Pathogens (AD[14]),TOS Science Design for Small Mammal Abundance and Diversity (AD[15]), and NEON Aquatic Sample Strategy Document (AD[16]).

3.3 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the following files:

- NEON Raw Data Validation for Ground beetles sampled from pitfall traps (NEON.DP0.10022) (AD[04])
- NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[05])
- NEON Raw Data Validation for TOS Small Mammal Abundance and Diversity (NEON.DP0.10001) (AD[06])
- NEON Raw Data Validation for Fish electrofishing, gill netting, and/or fyke netting counts (NEON.DP0.20107) (AD[07])

All variables reported in the published data (L1 data) are also provided separately in the files:

- NEON Data Variables for Ground beetle sequences DNA barcode (NEON.DP1.10020) (AD[08])
- NEON Data Variables for Mosquito sequences DNA barcode (NEON.DP1.10038) (AD[09])
- NEON Data Variables for Small mammal sequences DNA barcode (NEON.DP1.10076.001) (AD[10])
- NEON Data Variables for Fish sequences DNA barcode (NEON.DP1.20105.001) (AD[11])

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 Terrestrial Observation System (TOS) spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and Earth Gravitational Model 96 (EGM96) 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.

3.4 Temporal Resolution and Extent

3.4.1 Ground beetle sequences DNA barcode

The finest temporal resolution at which beetle DNA barcoding data will be tracked is trapping bout, a ~14-day interval during which pitfall traps are deployed. The collectDate (indicating when the trap was collected) and collectionDateAccuracy (derived from days each trap was deployed) are provided for each barcoded specimen. Additional details about sampling bout frequency can be found in the TOS Protocol and Procedure: Ground Beetle Sampling (AD[17]).

3.4.2 Mosquito sequences DNA barcode

The finest temporal resolution at which mosquito DNA barcoding data will be tracked is collection from a given trap, a 8-16 hour interval during which CO2 traps are deployed. The collectDate (indicating the date the trap was collected), the eventTime (indicating the time the trap was collected) and collectionDateAccuracy (derived from hours each trap was deployed; rounded up to the nearest day) are provided for each barcoded specimen. Additional details about sampling bout frequency can be found in the TOS Protocol and Procedure: Mosquito Sampling (AD[18]).

3.4.3 Small mammal sequences DNA barcode

The finest temporal resolution at which mammal DNA barcoding data will be tracked is trapping night, a ~12 hour interval during which mammal box traps are deployed. The collectDate (indicating the date the trap was collected) and collectionDateAccuracy (derived from hours each trap was deployed; rounded up to the nearest day) are provided for each barcoded specimen. Additional details about sampling bout frequency can be found in the TOS Protocol and Procedure: Small Mammal Sampling (AD[19]).

3.4.4 Fish sequences DNA barcode

The finest temporal resolution at which fish DNA barcoding data will be tracked is sampling pass (which may be electrofishing or gill-netting, or fyke netting), the interval during which fish are collected. The collectDate (indicating the date fish were collected) and collectionDateAccuracy (derived from the duration of the sampling; rounded up to the nearest day) are provided for each barcoded specimen. Additional details about sampling bout frequency can be found in the AOS Protocol and Procedure: Fish Sampling in Wadeable Streams (AD[20]) and AOS Protocol and Procedure: Fish Sampling In Lakes (AD[21]).

3.5 Spatial Resolution and Extent

Beetle traps are located within distributed base plots, mosquito traps are located at mosquito points, mammal traps are placed within mammal grids and fish are collected within a reach. Although finer scale spatial information is available from the field collection data associated with each of these data types, all DNA barcoding products use plot centroid information, which is coarser in resolution. Thus, the provided geographic coordinates will be offset from the collection location as follows:

- ~20 meters from the actual collection trap in the Ground beetle sequences DNA barcode product
- 0-10 meters from the actual collection trap in the Mosquito sequences DNA barcode product
- 0-64 meters from the actual collection trap in the Small mammal sequences DNA barcode product
- 0-750 meters from the actual collection location in the Fish sequences DNA barcode product

These uncertainties are reflected in the coordinateAccuracy values in the downloaded data.

3.6 Associated Data Streams

Each barcoded tissue specimen is referrenced on BOLD by a NEON-assigned **sampleID**. These sample identifiers are also present in other NEON data products, but are refered to under different variable names.

3.6.1 Ground beetle sequences DNA barcode

individualID is the linking variable that tie specific samples and associated metadata between the Ground beetles sampled from pitfall traps (NEON.DP1.10022) and Ground beetle sequences DNA barcode (NEON.DP1.10020). This identifier is present in the bet_parataxonomistID, bet_expertTaxonomistIDProcessed and bet_expertTaxonomistIDRaw tables within the column **individualID**.

3.6.2 Mosquito sequences DNA barcode

individualID is the linking variable that tie specific samples and associated metadata between the Mosquitoes sampled from CO2 traps (NEON.DP1.10043) and Mosquito sequences DNA barcode (NEON.DP1.10038) data products. This identifier is present in the mos_identification table within the individualID of each pinned individual is given in a pipe-delimited list in the column labelled **individualIDList**.

3.6.3 Small mammal sequences DNA barcode

earSampleID is a linking variable that ties specific samples and associated metadata from the Small Mammal Box Trapping (NEON.DP1.10072) to the Small mammal sequences DNA barcode data product (NEON.DP1.10076). This identifier is present in the mam_pertrapnight table within the column **earSampleID**.

3.6.4 Fish sequences DNA barcode

dnaSampleID is a linking variable that ties specific samples and associated metadata from the Fish electrofishing, gill netting, and/or fyke netting counts to the Fish sequences DNA barcode data product. This identifier is present in the fsh_perFish table within the column **dnaSampleID**.

3.7 Product Instances

3.7.1 Ground beetle sequences DNA barcode

No more than 95 individuals per site per year will receive DNA barcoding.

3.7.2 Mosquito sequences DNA barcode

No more than 95 individuals per domain per year will receive DNA barcoding.

3.7.3 Small mammal sequences DNA barcode

No more than 95 individuals per domain per year will receive DNA barcoding.

3.7.4 Fish sequences DNA barcode

No more than 95 individuals per domain per year will receive DNA barcoding.

3.8 Data Relationships

3.8.1 Ground beetle sequences DNA barcode

See the NEON User Guide to Ground beetles sampled from pitfall traps (NEON.DP1.10022) for details about beetle collection and data relationships. Duplicates and/or missing data may exist where protocol and/or data entry abberations have occurred; users should check data carefully for anomalies before joining tables.

bet_BOLDvoucherInfo.csv - > One record expected per sampleID for all time; max of 95 records per site per year.

bet_BOLDtaxonomy.csv - > One record expected per sampleID for all time; max of 95 records per site per year.

bet_BOLDspecimenDetails.csv - > One record expected per sampleID for all time; max of 95 records per site per year.

bet_BOLDcollectionData.csv - > One record expected per sampleID for all time; max of 95 records per site per year.

3.8.2 Mosquito sequences DNA barcode

See the NEON User Guide to Mosquitoes sampled from CO2 traps (NEON.DP1.10043) and Mosquito-borne pathogen status (NEON.DP1.10041) for details about mosquito collection and data relationships. Duplicates and/or missing data may exist where protocol and/or data entry abberations have occurred; users should check data carefully for anomalies before joining tables.

mos_BOLDvoucherInfo.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mos_BOLDtaxonomy.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mos_BOLDspecimenDetails.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mos_BOLDcollectionData.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

3.8.3 Small mammal sequences DNA barcode

See NEON User Guide to Small Mammal Box Trapping (NEON.DP1.10072) for details about small mammal collection and data relationships. Duplicates and/or missing data may exist where protocol and/or data entry abberations have occurred; users should check data carefully for anomalies before joining tables.

mam_BOLDvoucherInfo.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mam_BOLDtaxonomy.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mam_BOLDspecimenDetails.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

mam_BOLDcollectionData.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

3.8.4 Fish sequences DNA barcode

See NEON User Guide to Fish electrofishing, gill netting, and/or fyke netting counts for details about fish collection and data relationships. Duplicates and/or missing data may exist where protocol and/or data entry abberations

have occurred; users should check data carefully for anomalies before joining tables.

fsh_BOLDvoucherInfo.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

fsh_BOLDtaxonomy.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

fsh_BOLDspecimenDetails.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

fsh_BOLDcollectionData.csv - > One record expected per sampleID for all time; max of 95 records per domain per year.

3.9 Special Considerations

For ease of integration with external datasets, DNA barcoding data (along with other assembled resources for each specimen) are published on the Barcode of Life Datasystem (BOLD, http://www.barcodinglife.com/). Note that many early (pre-2014) NEON specimens for submitted to BOLD were collected during field prototype campaigns or obtained from museum archives (Gibson et al. 2012) and may not correspond to production data available through NEON's data portal.

In some cases, parataxonomist or taxonomist identification of speciments will conflict with the taxonomic assignment based on DNA sequence data. Data users are encouraged to consider a number of factors in resolving conflicting identifications, including the physical condition of the specimen, the length and quality of the DNA sequence, the phylogenetic clarity of the taxonomic designation provided by the taxonomist and DNA barcoding.

3.9.1 Retrieving DNA barcoding sequence data from BOLD

There are a number of ways to search and retrieve DNA barcoding data. Here, the workflow for finding beetle DNA sequence data is described, but the process applies to all four DNA sequence products.

1. From the NEON data portal: Under the 'Links to Data' header, the link "BOLD Project: Ground beetle sequences DNA barcode" will take the user to boldsystems.org website and return sequence data for all locations queried by the user on the NEON data portal. For example, if 'CPER' and 'WOOD' are selected on the NEON data portal, all Ground Beetle DNA sequences on BOLD collected from the Central Plains Experimental Range (CPER) and Woodworth (WOOD) will be displayed. This is a dynamic link that will automatically update based on the sites and data ranges provided in the user query.
2. From BOLD directly: Users who are interested in using the BOLD data analysis pipeline may want to combine NEON datasets with other datasets. This may be more easily achieved by querying the BOLD database directly. All DNA sequence data provided by NEON may be searched by querying 'National Ecological Observatory Network, United States' as the institution. Through the BOLD website, users can analyze samples from a variety of NEON and non-NEON projects.

About the Barcode of Life Database

All NEON DNA sequence data are provided under the campaign moniker 'National Ecological Observatory Network DNA Barcoding'. All operational sequence data are categorized into four projects under this umbrella campaign.

- Ground beetle sequences DNA barcode (project code: BETN)
- Mosquito sequences DNA barcode (project code: MOSN)
- Small mammal sequences DNA barcode (project code: MAMN)
- Fish sequences DNA barcode (project code: FSHN)

For the 30-year life of the observatory, all new operational data will be stored in one of the above 4 campaigns (corresponding to its data type). Additional prototype DNA sequence data may be found in other projects under the 'National Ecological Observatory Network DNA Barcoding' campaign.

4 DATA QUALITY

4.1 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 following documents, provided with download of the pertinent data product:

- NEON Raw Data Validation for Ground beetles sampled from pitfall traps (NEON.DP0.10022) (AD[04])
- NEON Raw Data Validation for Mosquitoes sampled from CO2 traps (NEON.DP0.10043) (AD[05])
- NEON Raw Data Validation for TOS Small Mammal Abundance and Diversity (NEON.DP0.10001) (AD[06])
- NEON Raw Data Validation for Fish electrofishing, gill netting, and/or fyke netting counts (NEON.DP0.20107) (AD[07])

Contained within each file is a field named 'entryValidationRulesForm', which describes syntactically the validation rules for each field built into the data entry application. Also included in this file is a field named 'entryValidationRulesParser', which describes syntactically the validation rules for each field that is performed upon ingest of the data into the NEON Cyberinfrastructure, based on a standardized data validation language (Nicl) internal to NEON. Please see AD[25] for more information about the Nicl language.

Data collected prior to 2017 were processed using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow.

4.2 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[24]).

4.3 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.

4.4 Quality Flagging

Per BOLD requirements, no quality flagging of sequence metadata is provided on the NEON portal.

4.5 Analytical Facility Data Quality

DNA sequence data is provided on the Barcode of Life website for each tissue sample analyzed. Sequence data are provided in two ways: as a consensus sequence (available as a fasta file) and as a batch of trace files originating from the sequencer (a zip of all generated ab1 files). COI sequences are reviewed for quality by the analytical facility & the BOLD platform. High quality sequences (>507 base pairs, less than 1% ambiguous bases, and no stop codon or contamination flags) are cross-searchable on the BOLD Public Data Portal and Barcode Index Numbers (BINs) Database. Tissue samples that yield low quality sequence are flagged as problematic; see below for an explanation of common quality flags specific to BOLD. Low quality sequence are not assigned BINs, but are still searchable on the BOLD Public Data Portal.

Quality Flag	definition
Stop Codon	DNA sequence contained a stop codon. COI is a protein-coding gene, and thus should not include stop codons (this would disrupt the translation of the DNA sequence into amino acids). Typically indicative that primers amplified only part of the COI locus or a psuedo-gene
Contamination	Double peaks at nearly every base position. If two or more sequences from unrelated species are co-amplified, this contamination will yield a trace file that cannot be interpreted. Consensus sequence will have a large number of ambiguous base pair assignments.

5 REFERENCES

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3:294–299.

Gibson, C. M., R. H. Kao, K. K. Blevins, and P. D. Travers. 2012. Integrative taxonomy for continental-scale terrestrial insect observations. PLoS ONE 7:e37528.

Hebert, P., A. Cywinska, S. Ball, and J. DeWaard. 2003. Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings, Biological sciences/The Royal Society, 270(1251), 313–321. Proceedings of the Royal Society B-Biological Sciences 270:313–321.