NEON USER GUIDE TO MACROINVERTEBRATE METABARCODING (DP1.20126.001) AND ZOOPLANKTON METABARCODING (DP1.20221.001)

PREPARED BY	ORGANIZATION
Stephanie Parker	AQU
Lee Stanish	FSU

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	3/07/2018	Initial Release
B	10/20/2020	Changing DP names to 'Macroinvertebarate metabarcoding' and 'Zooplankton metabarcoding'; including raw and bioinformatics data tables and data relationships. Included general statement about usage of neonUtilities R package and statement about possible location changes. Updated taxonomy information. Updated taxonomy information. Updated littoral sampling locations and figures.
C	03/02/2022	Remove references to expanded data package, note that all tables are available in the basic package. Clarify taxon table use. Added language in Section 4 Taxonomy addressing RTE species obfuscation in the data. Updated section 5.4 Data Revision with latest information regarding data release.
C.1	03/23/2022	Clarify field sample relationship with morphological samples and individualCount in metabarcodeTaxonomy tables.
D	12/08/2022	Add metabarcodeTaxonomyStandard table and remove zooplankton L0 ingest, all data are ingested in DP0.20126.001.
D.1	1/10/2024	Add dataQF description for low extraction efficiency
E	04/17/2024	Minor formatting updates
E.1	10/10/2024	Adding data QF for macroinvertebrate kicknet

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 macroinvertebrate or zooplankton DNA samples collected in the field 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 products Macroinvertebrate metabarcoding and Zooplankton metabarcoding 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 files NEON Data Variables for Macroinvertebrate metabarcoding (DP1.20126.001) (AD[04]) and NEON Data Variables for Zooplankton metabarcoding (DP1.20221.001) (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 AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling (AD[07]) and AOS Protocol and Procedure: Zooplankton Sampling in Lakes (AD[08]). The raw data that are processed in this document are detailed in the files NEON Raw Data Validation for Macroinvertebrate metabarcoding (DP0.20126.001) (AD[03]), 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., '20221') as the corresponding L1 data product.

2 RELATED DOCUMENTS

2.1 Associated Documents

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.002652	NEON Data Products Catalog
AD[03]	Available with data download	Macroinvertebrate Validation csv
AD[04]	Available with data download	Macroinvertebrate Variables csv
AD[05]	Available with data download	Zooplankton Variables csv
AD[06]	NEON.DOC.001152	NEON Aquatic Sampling Strategy
AD[07]	NEON.DOC.003046	AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling
AD[08]	NEON.DOC.001194	AOS Protocol and Procedure: Zooplankton Sampling in Lakes
AD[09]	NEON.DOC.000008	NEON Acronym List
AD[10]	NEON.DOC.000243	NEON Glossary of Terms
AD[11]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[12]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
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 metabarcoding data products provide DNA sequence data and metadata for macroinvertebrate and zooplankton communities at NEON aquatic sites. These data may be used for determination of diversity patterns in bulk samples, and analogous morphological taxonomy samples for both macroinvertebrates and zooplankton are collected at the same time and location, so taxonomic data may be correlated by data users. NEON uses PCR amplification of two target regions of the CO1 gene using primers described in Gibson et al. (2015). The use of two distinct primer sets and regions enables greater coverage of the diversity of distinct arthropod groups. Sequence data are generated using high-throughput technology that produces many thousands of sequence reads per sample (Armougom and Didier 2009, Klindworth et al. 2013).

The type of sampler used to collect samples in the field is determined by the habitat and substrate type (macroinvertebrates) or water depth (zooplankton) at the sampling location. Macroinvertebrates are collected using a Surber sampler, modified kicknet, hand corer, or D-frame net (used in lakes and rivers only). All sampling devices collect material from a known area of the benthos. For zooplankton samples, locations deeper than 4 m are sampled using a vertical tow net, while locations shallower than 4 m are sampled using a Schindler-Patalas sampler (USEPA 2012a, 2012b). Typically, multiple (up to 3) tows or Schindler traps are collected and composited into a single sample. Zooplankton samples are collected on a volumetric basis.

Sample collection methods differ between macroinvertebrate and zooplankton samples, but in general samples are minimally processed in the field in order to reduce the introduction of microbial contaminants. After collection, samples are preserved in 95% ethanol and chilled at 4 degrees C, with an ethanol change occurring within 24 hours of collection. For additional information see sampling design NEON Aquatic Sampling Strategy (AD[06]), and protocols AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling (AD[07]) and AOS Protocol and Procedure: Zooplankton Sampling in Lakes (AD[08]). Samples are shipped to an sequencing laboratory where sample homogenization, DNA extraction, sequencing library preparation, DNA sequencing, and bioinformatics analysis occur.

In 2018, a mock macroinvertebrate sample was created as a standard to be run in each sequencing run. This standard samples uses morphologically identified specimens from the Macroinvertebrate collection data product (DP1.20120.001). A list of organisms composing the standard sample can be found the inv_dnaStandardTaxon table. The standard sample is preserved in 95% ethanol and shipped to the analytical laboratory, where it is homogenized and DNA is extracted. Aliquots of the DNA extract are frozen and stored at the archive facility for future use. Extracted DNA from the mock sample is included the each sampling run starting in 2019. When aliquots of the mock sample are used up, NEON will create a new mock sample.

3.1 Spatial Sampling Design

Benthic invertebrates at NEON aquatic sites (Figure 1) are sampled using a percent-based macrohabitat approach (after Moulton et al. 2002). Habitats sampled focus on riffles, runs, pools, and step pools depending on the percent cover of each habitat within each 1 km-long NEON Aquatic wadeable stream site (NOTE: some NEON sites may be less than 1 km due to permitting restrictions), and benthic-pelagic and

littoral samples in lakes and rivers. Three samples are collected in the dominant habitat type (wadeable stream) or littoral area (lake and river) on a given sampling date at a site.

Samplers used for macroinvertebrate collection are designed to work by disturbing the benthic sediments and catching invertebrates in an attached net or container, while delineating the benthic area sampled for a quantitative result. The sampler type chosen differs depending on the water depth, velocity, and substratum type in the chosen habitat (Hauer and Resh 2006). The collection method may differ depending on the habitat and substrate being sampled, however all samples are collected from the surface of the natural substratum in each habitat using a quantitative sampling method. See AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling (AD[07]) for additional details on sampling strategy and SOPs.

Zooplankton samples are collected only from lakes (Figure 2). Samples are collected near the littoral sensors and buoy (deepest) sampling locations, and are designed to sample organisms inhabiting the water column. The type of sampler selected depends on the depth of the lake at the sampling location, and the volume of lake water sampled can be calculated using the number of tows/traps and the volume of the sampler used. See AOS Protocol and Procedure: Zooplankton Sampling in Lakes (AD[08]) for additional details on sampling strategy and SOPs.

As much as possible, sampling occurs in the same locations over the lifetime of the Observatory. However, over time some sampling 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/

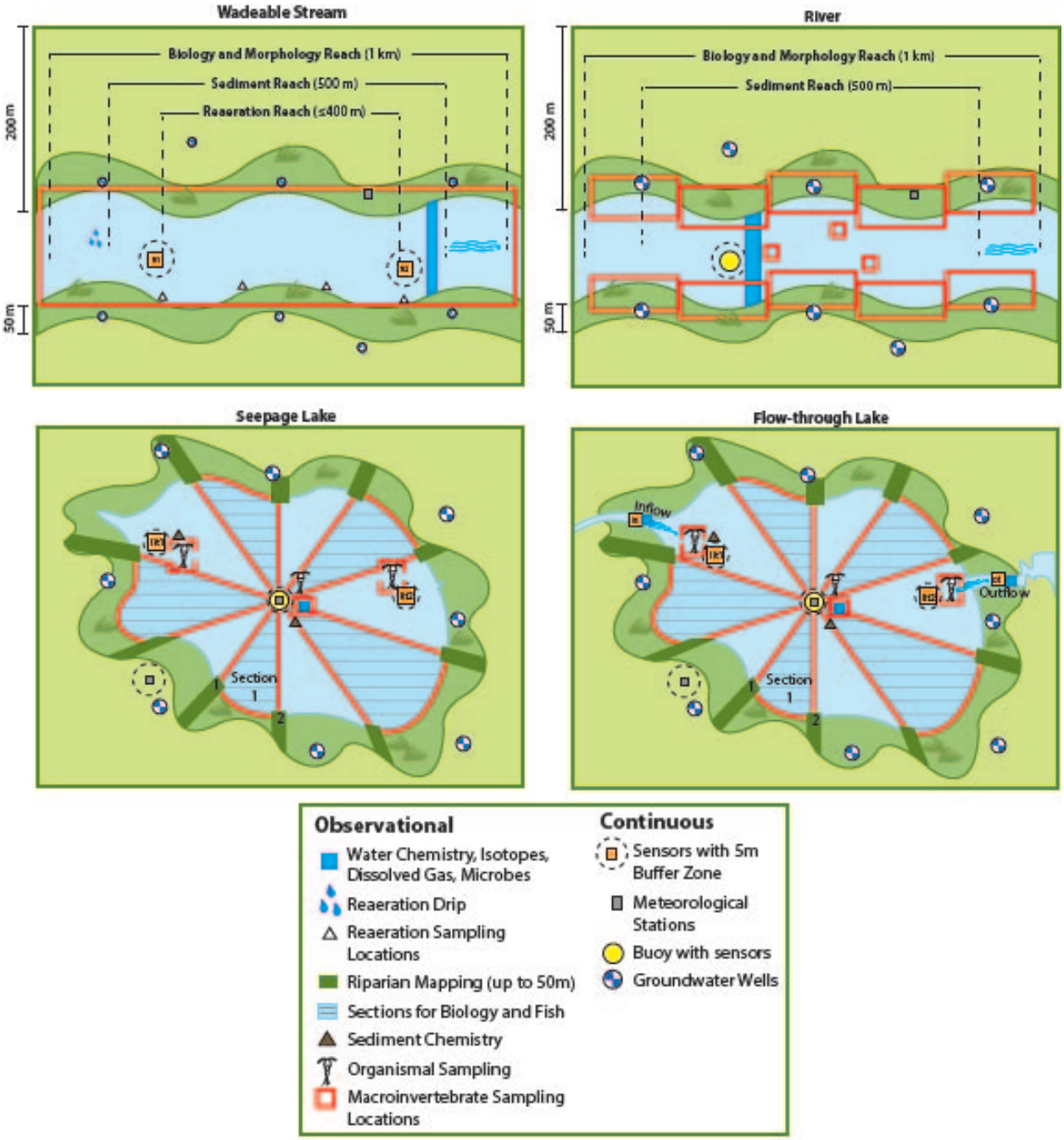

Figure 1: Generic aquatic site layout for lakes, river and wadeable streams, with macroinvertebrate sampling locations in red.

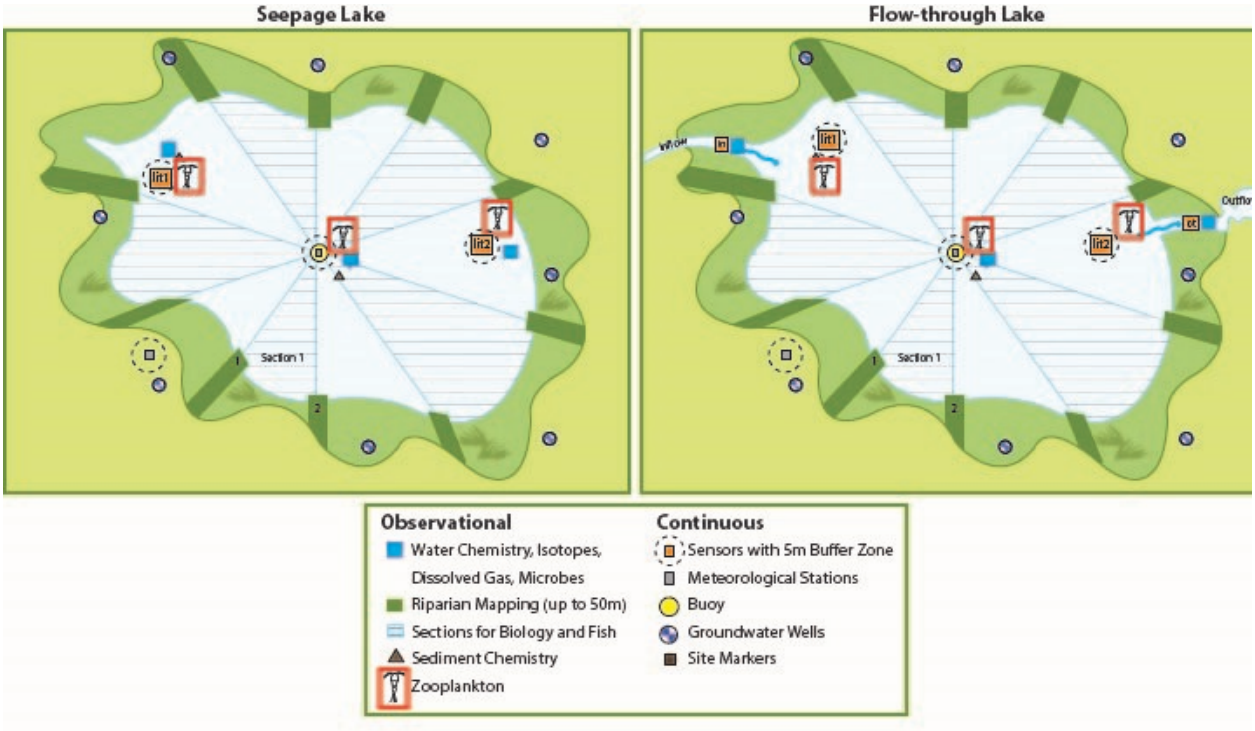

Figure 2: Generic aquatic site layout for lakes with zooplankton sampling locations in red.

3.2 Temporal Sampling Design

Sampling occurs three times per year at each NEON site (AD[06]). Timing of sampling is site-specific and determined based on historical hydrological and meteorological data. Sequencing analysis is done only on samples collected during Bout 2 (midsummer), while samples from Bouts 1 and 3 (spring and autumn) are preserved and sent to the NEON Biorepository for archiving and use by the external community.

Sample Bout 1 is an early-season date, representing a period of rapid biomass accumulation after winter, typically prior to leaf out or after ice-off where applicable. Sample Bout 2 targets mid-summer baseflow conditions and sample Bout 3 represents the late growing season (typically autumn) during leaf-fall where applicable. These dates differ on a site-by-site basis, but should always occur at, or near, baseflow conditions within the watershed. Sampling does not occur directly following a rain or wind even that causes turbidity in the water column (lakes/rivers) or a flood in wadeable streams (defined as >1.5 x base flow; Biggs et al. 1999). Should such a flood event occur on or prior to a target collection date, sampling is delayed 3 days-1 week (maximum 2 weeks, dependent on field schedule) to allow for invertebrates to recolonize the substratum (c.f. Brooks and Boulton 1991, Matthaei et al. 1996). Sampling at each site is completed within a single day for each bout. See NEON Aquatic Sampling Strategy (AD[06]), AOS Protocol and Procedure: Aquatic Macroinvertebrate Sampling (AD[07]), and AOS Protocol and Procedure: Zooplankton Sampling in Lakes (AD[08]) for additional details.

3.3 Sampling Design Changes

- 2017: Data were collected during the fall sampling bout at D02 POSE (macroinvertebrates), D03 BARC (macroinvertebrates and zooplankton), and D03 SUGG (zooplankton) as a pilot study.
- 2018-present: Full sampling for 3 bouts per site per year. Samples from Bout 2 (mid-summer) are sent to an external facility for analysis. Samples from Bouts 1 (spring) and 3 (fall) are stored for use at the archive.
- 2019-present: A standard sample of known and identified macroinvertebrates was created to use as a positive control in all sequencing runs.
- 2020-present: Location names for the nearshore sensors in seepage lakes (lakes without a true inlet and outlet stream) were changed on January 1, 2021. The location previously known “inlet” changed to “littoral 1” (“lit2”) and “outlet” changed to “littoral 2” (“lit2”) to indicate that these locations are not near an inlet or outlet stream. Flow-through lakes (e.g., D18 TOOK) have sensors in the inflow and outflow streams, as well as the lit1 and lit2 locations in the lake.

3.4 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the files NEON Raw Data Validation for Macroinvertebrate metabarcoding (DP0.20126.001) (AD[03]). All variables reported in the published data (L1 data) are also provided separately in the files NEON Data Variables for Macroinvertebrate metabarcoding (DP1.20126.001) (AD[04]) and NEON Data Variables for Zooplankton metabarcoding (DP1.20221.001) (AD[05]).

Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 7 December 2017), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 7 December 2017), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 7 December 2017), where applicable. NEON Aquatic Observation System (AOS) 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.

3.5 Temporal Resolution and Extent

The finest temporal resolution that macroinvertebrate and zooplankton metabarcoding data will be tracked is per sampling day. All 3 samples per module (macroinvertebrate or zooplankton) are collected within a single day at a particular site. A suite of other biological sampling occurs at the site during the same ~30 day bout. Three sampling bouts occur per site per year.

The finest resolution at which temporal data are reported is at **collectDate**, the date and time of day when the samples were collected in the field.

The NEON Data Portal provides data in monthly files for query and download efficiency. Queries including any part of a month will return data from the entire month. Code to stack files across months is available here: https://github.com/NEONScience/NEON-utilities.

3.6 Spatial Resolution and Extent

Each macroinvertebrate sample represents a patch of stream bottom within the 1 km permitted wadeable or river reach, or permitted lake area, and contains multiple individuals. The exact location (latitude and longitude) of each sample is not tracked as it is intended to represent the overall habitat. The **namedLocation** reported in wadeable streams represents a midpoint in the permitted reach, plus coordinate uncertainty surrounding that point. In lakes and rivers, some samples are collected near monumented locations associated with a more-specific **namedLocation** (e.g., NEON sensor infrastructure). Sampling locations are tracked by latitude and longitude and include an indication of **coordinateUncertainty**.

Each zooplankton sample represents a location in a lake near one of the NEON sensor installations (littoral1, littoral2, or buoy), and contains multiple individuals. The **namedLocation** reported represents the location of the NEON sensor infrastructure near the sampling location, plus coordinate uncertainty surrounding that point. The protocol dictates that samples are collected approximately 5 m from the sensor infrastructure to minimize effects on the sensor data, so the standard **coordinateUncertainty** is 10 m to represent the normal sampling distance from the sampling location. If, for some reason, sampling cannot occur within 10 m of the named location, technicians will enter **additionalCoordinateUncertainty**.

Samples are collected from the dominant habitat type (wadeable streams), benthic littoral zone (lakes/rivers), or pelagic water column (lakes - zooplankton). Overall, this results in a spatial hierarchy of:

namedLocation (finest spatial resolution, ID of location within site) -> siteID (ID of NEON site) -> domainID (ID of a NEON domain)

3.7 Associated Data Streams

Macroinvertebrate and zooplankton metabarcoding samples are collected at the same time and location, and using the same method, as an analogous morphological taxonomy sample. Metabarcoding and morphological field samples are collected as separate field samples in the same habitat unit (e.g., same riffle) and one is not a subsample of the other. Related samples share the same **eventID** and **namedLocation**, as well as the same root **sampleID**. Metabarcoding sample IDs are equal to the taxonomic sampleID + "DNA" appended to the end. Taxonomic data are available in the NEON data products "Macroinvertebrate collection" (DP1.20120.001) and "Zooplankton collection" (DP1.20219.001).

3.8 Product Instances

At each aquatic site, there will be up to 3 macroinvertebrate samples collected and sequenced and 6 samples collected and sent to the NEON Biorepository per year (3 macroinvertebrate samples per sampling

bout). At lake sites, there will be up to 3 zooplankton samples collected and sequenced and 6 samples collected and sent to the NEON Biorepository per year. Each sample that is sequenced may generate multiple records from the external lab. Sampling impractical records are created in instances where samples were not able to be collected during the expected sampling bout.

3.9 Data Relationships

3.9.1 Macroinvertebrate metabarcoding (DP1.20120.001)

For each macroinvertebrate sampling event, a record is created in inv_fieldData. All macroinvertebrate taxonomic samples will appear here along with **geneticSampleID**, whether a DNA sample was collected or not (unlike the zooplankton data product, where only **geneticSampleID** will appear in the table). In the event that sampling is impractical (e.g., the location is dry, ice covered, etc.) or no **geneticSampleID** is taken, and there will be no child records. Otherwise, there may be a number of child records in subsequent tables. Child records will be found in inv_dnaExtraction (initial subsampling and dna extraction metadata at the external facility), inv_pcrAmplification (PCR metadata), inv_markerGeneSequencing (sequencing metadata), inv_dnaRawDataFiles (unprocessed data available for download), and inv_metabarcodeTaxonomy (bioinformatics data). Every record in inv_dnaExtraction, inv_pcrAmplification, inv_markerGeneSequencing, and inv_metabarcodeTaxonomy has a corresponding record in inv_fieldData describing field collection conditions, location, and metadata during sample collection. The **dnaSampleID** is created in the inv_dnaExtraction table, linking downstream data to the **geneticSampleID** from the inv_fieldData table. There is one unique record for each **dnaSampleID** in inv_dnaExtraction unless extractions were unsuccessful and multiple extractions were required, while inv_pcrAmplification, inv_markerGeneSequencing, and inv_dnaRawDataFiles may have multiple records per **dnaSampleID**, corresponding to different forward and reverse primer sets. The bioinformatics table (inv_metabarcodeTaxonomy) will have multiple records per **dnaSampleID** corresponding to each taxon sequence detected in the analysis. Duplicates and/or missing data may exist where protocol and/or data entry aberrations have occurred; users should check data carefully for anomalies before joining tables.

An aliquot of the standard (mock) invertebrate sample, created in a NEON lab, is included with every sampling run. Information on the composition of the mock sample is available in the table inv_dnaStandardTaxon, which can be linked to the inv_dnaExtractionStandard table using the **dnaStandardSampleID**. Each **dnaStandardSampleID** is subset into several aliquots. Each aliquot receives a unique **dnaSampleID** in the inv_dnaExtractionStandard table. Mock sample data are organized in the same was as the data above, starting with the inv_dnaExtractionStandard table. There is one unique record for each **dnaSampleID** in inv_dnaExtraction unless extractions were unsuccessful and multiple extractions were required, while inv_pcrAmplification, inv_markerGeneSequencing, and inv_dnaRawDataFiles may have multiple records per **dnaSampleID**, corresponding to different **replicate**s.

inv_fieldData.csv -> One record is created for each sample collected in the field, creating a unique **geneticSampleID** representing one sample per **collectDate** and **namedLocation**. This table also indicates field conditions, including **samplerType**, **habitatType**, and **benthicArea**.

inv_dnaExtraction.csv -> The **geneticSampleID** from the fieldData table is subsampled to create the **dnaSampleID** in the inv_dnaExtraction table. Generally, there will be only one DNA extraction per

dnaSampleID, but in some cases multiple extractions will be necessary.

inv_pcrAmplification.csv -> Metadata on PCR amplification sample processing is presented in this table. One record is expected per **dnaSampleID**/**replicate** combination. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_markerGeneSequencing.csv -> Metadata on gene sequencing is presented in this table. One record is expected per **dnaSampleID**/**replicate** combination. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_dnaRawDataFiles.csv -> Metadata directing a user to the raw data file URL is presented in this table in the field **rawDataFilePath**. Raw data files are presented in a zipped format. One record is expected in this table per **dnaSampleID**/**rawDataFileName** combination, this results in 4 records per **dnaSampleID**: 1 forward and 1 reverse sequence for the BE primer, and 1 forward and 1 reverse sequence for the F230 primer. There are 4 raw files per year in the url location: 1 forward BE, 1 reverse BE, 1 forward F230, and 1 reverse F230 containing. Each file contains sequences from all **dnaSampleID**s from that sampling year. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_metabarcodeTaxonomy.csv -> Bioinformatics data per **dnaSampleID** is presented in this table. Several records are expected per **dnaSampleID**, one for each taxon reported with the sequence or read count in the **individualCount** field. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_metabarcodeTaxonList.csv -> The taxon library at the time of the bioinformatics processing is reported approximately once per year, and the URL for each year/upload is reported with the URL in this table. The URL directs a user to a large file containing the entire taxonomic library used in the bioinformatics pipeline. Records are linked to the inv_metabarcodeTaxonomy via the **sequenceName** and **processedDate**.

inv_metabarcodeSummary.csv -> Metadata describing the bioinformatics pipeline and referenceDatabase is reported in this table. Data may be related to precious table using the **processedDate** that falls within the range of the **labSpecificStartDate** to the **labSpecificEndDate**.

inv_dnaExtractionStandard.csv -> The **geneticSampleID** from the dnaStandard field table is subsampled to create the **dnaSampleID** in the inv_dnaExtractionStandard table. Generally, there will be only one DNA extraction per **dnaSampleID**, but in some cases multiple extractions will be necessary.

inv_pcrAmplificationStandard.csv -> Metadata on PCR amplification sample processing for the mock sample is presented in this table. One record is expected per **dnaSampleID**. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table.

inv_markerGeneSequencingStandard.csv -> Metadata on gene sequencing for the mock sample is presented in this table. One record is expected per **dnaSampleID**. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table.

inv_dnaRawDataFilesStandard.csv -> Metadata directing a user to the raw data file URL for the mock sample is presented in this table in the field **rawDataFilePath**. Raw data files are presented in a zipped format. One record is expected in this table per **dnaSampleID**/**rawDataFileName** combination, this results in 4 records per **dnaSampleID**: 1 forward and 1 reverse sequence for the BE primer, and 1 forward and 1 reverse sequence for the F230 primer. There are 4 raw files per year in the url location: 1 forward BE, 1

reverse BE, 1 forward F230, and 1 reverse F230 containing. Each file contains sequences from all **dnaSampleID**s from that sampling year. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_metabarcodeTaxonomyStandard.csv -> Bioinformatics data per **dnaSampleID** from the mock sample is presented in this table. Several records are expected per **dnaSampleID**, one for each taxon reported with the sequence or read count in the **individualCount** field. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table.

inv_dnaStandardTaxon.csv -> The inv_dnaStandardTaxon table documents the **scientificName**, **sizeClass**, and **individualCount** of the organisms added to the dna mock sample. Each record points to the parent data in the inv_perTaxon.csv data from the Macroinvertebrate Collection data product (DP1.20120.001), and assigns it to the mock sample through the **dnaStandardSampleID** in inv_dnaExtractionStandard.

3.9.2 Zooplankton metabarcoding (DP1.20221.001)

For each zooplankton sampling event where a genetic sample is collected, a record is created in zoo_fieldData with a unique **geneticSampleID**.In the event that sampling is impractical (e.g., the location is dry, ice covered, etc.) or no **geneticSampleID** is taken, no zoo_fieldData records will appear in the download. If sampling occurs, there may be a number of child records in subsequent tables. Child records will be found in zoo_dnaExtraction (initial subsampling and dna extraction metadata at the external facility), zoo_pcrAmplification (PCR metadata), zoo_markerGeneSequencing (sequencing metadata), and zoo_metabarcodeTaxonomy (bioinformatics data). Every record in zoo_dnaExtraction, zoo_pcrAmplification, zoo_markerGeneSequencing, and zoo_metabarcodeTaxonomy has a corresponding record in zoo_fieldData describing field collection conditions, location, and metadata during sample collection. The **dnaSampleID** is created in the zoo_dnaExtraction table, linking to **geneticSampleID** from the zoo_fieldData table (zoo_dnaExtraction.geneticSampleID=zoo_fieldData.sampleID). There is one unique record for each **dnaSampleID** in zoo_dnaExtraction unless extractions were unsuccessful and multiple extractions were required, while zoo_pcrAmplification, zoo_markerGeneSequencing, and zoo_dnaRawDataFiles may have multiple records per **dnaSampleID**, corresponding to different forward and reverse primer sets. The bioinformatics table (zoo_metabarcodeTaxonomy) will have multiple records per **dnaSampleID** corresponding to each taxon sequence detected in the analysis. Duplicates and/or missing data may exist where protocol and/or data entry aberrations have occurred; users should check data carefully for anomalies before joining tables.

An aliquot of the standard (mock) invertebrate sample, created in a NEON lab, is included with every sampling run and copied from the Macroinvertebrate metabarcoding data product to the Zooplankton metabarcoding data product. Information on the composition of the mock sample is available in the table inv_dnaStandardTaxon, which can be linked to the inv_dnaExtractionStandard table using the **dnaStandardSampleID**. Each **dnaStandardSampleID** is subset into several aliquots. Each aliquot receives a unique **dnaSampleID** in the inv_dnaExtractionStandard table. Mock sample data are organized in the same was as the data above, starting with the inv_dnaExtractionStandard table. There is one unique record for each **dnaSampleID** in inv_dnaExtraction unless extractions were unsuccessful and multiple extractions were required, while inv_pcrAmplification, inv_markerGeneSequencing, and inv_dnaRawDataFiles may have multiple records per **dnaSampleID**, corresponding to different **replicate**s.

zoo_fieldData.csv - > One record is created for each sample collected in the field, creating a unique **geneticSampleID**. This table also indicates the field conditions, including **samplerType**, number of tows or traps collected (**towsTrapsNumber**), and sampling depth (**zooDepth1**, **zooDepth2**, **zooDepth3**).

zoo_dnaExtraction.csv - > The **geneticSampleID** from the fieldData table is subsampled to to create the **dnaSampleID** in the inv_dnaExtraction table. One record is expected per **dnaSampleID** here, and will be linked to subsequent tables. Generally, there will be only one DNA extraction per **dnaSampleID**, but in some cases multiple extractions will be necessary.

zoo_pcrAmplification.csv - > Metadata on PCR amplification sample processing is presented in this table. One record is expected per **dnaSampleID**/**replicate** combination. The **dnaSampleID** equals the **dnaSampleID** in the zoo_dnaExtraction table.

zoo_markerGeneSequencing.csv - > Metadata on gene sequencing is presented in this table One record is expected per **dnaSampleID**/**replicate** combination. The **dnaSampleID** equals the **dnaSampleID** in the zoo_dnaExtraction table.

zoo_dnaRawDataFiles.csv -> Metadata directing a user to the raw data file URL is presented in this table in the field **rawDataFilePath**. Raw data files are presented in a zipped format. One record is expected in this table per **dnaSampleID**/**rawDataFileName** combination, this results in 4 records per **dnaSampleID**: 1 forward and 1 reverse sequence for the BE primer, and 1 forward and 1 reverse sequence for the F230 primer. There are 4 raw files per year in the url location: 1 forward BE, 1 reverse BE, 1 forward F230, and 1 reverse F230 containing. Each file contains sequences from all **dnaSampleID**s from that sampling year. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

zoo_metabarcodeTaxonomy.csv -> Bioinformatics data per **dnaSampleID** is presented in this table. Several records are expected per **dnaSampleID**, one for each taxon reported with the sequence or read count in the **individualCount** field. . The **dnaSampleID** equals the **dnaSampleID** in the zoo_dnaExtraction table.

The following standard tables are republished from the macroinvertebrate metabarcoding data product: inv_metabarcodeTaxonList.csv -> The taxon library at the time of the bioinformatics processing is reported approximately once per year, and the URL for each year/upload is reported with the URL in this table. The URL directs a user to a large file containing the entire taxonomic library used in the bioinformatics pipeline. Records are linked to the zoo_metabarcodeTaxonomy via the **sequenceName** and **processedDate**. This table is copied from the Macroinvertebrate metabarcoding data product.

inv_metabarcodeSummary.csv -> Metadata describing the bioinformatics pipeline and referenceDatabase is reported in this table. Data may be related to precious table using the **processedDate** that falls within the range of the **labSpecificStartDate** to the **labSpecificEndDate**. This table is copied from the Macroinvertebrate metabarcoding data product.

inv_dnaExtractionStandard.csv -> The **geneticSampleID** from the dnaStandard field table is subsampled to create the **dnaSampleID** in the inv_dnaExtractionStandard table. Generally, there will be only one DNA extraction per **dnaSampleID**, but in some cases multiple extractions will be necessary. This table is copied from the Macroinvertebrate metabarcoding data product.

inv_pcrAmplificationStandard.csv -> Metadata on PCR amplification sample processing for the mock sample is presented in this table. One record is expected per **dnaSampleID**. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table. This table is copied from the Macroinvertebrate metabarcoding data product.

inv_markerGeneSequencingStandard.csv -> Metadata on gene sequencing for the mock sample is presented in this table. One record is expected per **dnaSampleID**. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table. This table is copied from the Macroinvertebrate metabarcoding data product.

inv_dnaRawDataFilesStandard.csv -> Metadata directing a user to the raw data file URL for the mock sample is presented in this table in the field **rawDataFilePath**. Raw data files are presented in a zipped format. One record is expected in this table per **dnaSampleID**/**rawDataFileName** combination, this results in 4 records per **dnaSampleID**: 1 forward and 1 reverse sequence for the BE primer, and 1 forward and 1 reverse sequence for the F230 primer. There are 4 raw files per year in the url location: 1 forward BE, 1 reverse BE, 1 forward F230, and 1 reverse F230 containing. Each file contains sequences from all **dnaSampleID**s from that sampling year. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtraction table.

inv_metabarcodeTaxonomyStandard.csv -> Bioinformatics data per **dnaSampleID** from the mock sample is presented in this table. Several records are expected per **dnaSampleID**, one for each taxon reported with the sequence or read count in the **individualCount** field. The **dnaSampleID** equals the **dnaSampleID** in the inv_dnaExtractionStandard table.

Data downloaded from the NEON Data Portal are provided in separate data files for each site and month requested. All tables for this data product are in the "basic" download package. 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

3.10 Special Considerations

For ease of integration with external data sets, some metabarcode sequence data are published on public sequence repositories. The primary data repository is MG-RAST (http://metagenomics.anl.gov, Meyer et al., 2008), which directly synchronizes its data with the European Bioinformatics Institute (EMBL-EBI) database and, through EMBL-EBI, synchronizes with the National Center for Biotechnology Information's Sequence Read Archive (SRA). A suite of metadata, compliant with minimum metadata standards defined by the Genomics Standards Consortium (e.g. MIxS, MIMARKS), accompanies the sequence data. While efforts are made to publish comprehensive sequencing metadata with the sequence data stored at public sequence repositories, potentially important data will only be available through the NEON Data Portal. These data include:

- Methods and SOPs
- QA data
- Sample identifiers to enable joining metabarcoding data with other related Data Products
- Data for other related Data Products

The sequence data for each region of the CO1 gene amplified are uploaded separately, such that there will be one file per sample per gene region targeted.

3.10.1 Retrieving Metabarcode Sequence Data

There are a number of ways to search and retrieve minimally processed metabarcode sequence data.

- From the NEON data portal:
    1. Directly from the Macroinvertebrate metabarcoding and Zooplankton metabarcoding data products via links in inv_dnaRawData files and zoo_dnaRawData files
    2. Links beginning with “MG-RAST Project: NEON Macroinvertebrate metabarcoding” will take the user to the MG-RAST project page for the queried data. This is a dynamic link and will automatically update based on the user query.
    3. The link “MG-RAST Project: NEON Zooplankton metabarcoding” will take the user to the MG-RAST project page for the queried data. This is a dynamic link and will automatically update based on the user query.
    4. The link “MG-RAST Sample Search” takes the user to the MG-RAST page for searching individual records, pre-populated with NEON records based on the user query.
- From MG-RAST directly: Users who are interested in using the MG-RAST data analysis pipeline may want to combine NEON datasets with other datasets. This may be more easily achieved by querying the MG-RAST database directly. Users can analyze samples from a variety of NEON and non-NEON projects. A free user account may be required.
- From SRA directly: Data and metadata are available for download from the SRA using the SRA toolkit. Documentation on how to install and use the toolkit for downloading sequence data is available on the SRA website.
- From EMBL-EBI: MG-RAST also synchronizes data sets with the European Bioinformatics Initiative Repository (EMBL-EBI, https://www.ebi.ac.uk/), which has a web and API interface for downloading data. The NEON macroinvertebrate metabarcoding data can be found by querying the NCBI Project ID PRJNA391345, and the NEON zooplankton metabarcoding data can be found by querying the NCBI Project ID PRJNA391744.

Note: There may be lags between publication of metadata on the NEON data portal and availability of sequence data on the public sequence repository.

4 TAXONOMY

4.1 *inv_dnaStandardTaxon*

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 macroinvertebrates and zooplankton was originally based on comprehensive taxonomy lists provided by expert taxonomy labs (EcoAnalysts, Inc. and GEI Consultants, Inc.) that were cross-referenced with taxonomic concepts from the Integrated Taxonomic Information System (ITIS, itis.gov) or Catalogue of Life (www.catalogueoflife.org) databases. Unique Taxon ID codes used to identify taxonomic concepts in the NEON master taxonomy list were generated for each taxon by concatenating the first three letters of the genus name together with the first three letters of the specific epithet to make a unique taxon ID for each scientific name. The list includes a variety of macroinvertebrate taxa, including mollusks, snails, worms, insects, mites, and crustaceans. NEON plans to keep the taxonomy updated in accordance with Merritt et al. (2019) and other current literature starting in 2020 and annually thereafter. The NEON taxonomy table was only used to check taxa in the inv_dnaStandardTaxon where specimens added to the mock macroinvertebrate sample/community standard.

The master taxonomy list also indicates the expected geographic distribution for each species by NEON domain and whether it is known to be introduced or native in that part of the range. Given that the spatial distributions of many aquatic macroinvertebrate taxa are not well known, NEON assumes that all taxa are possible at all aquatic sites. As spatial resolution of distribution maps improves, NEON will update the taxon tables to generate errors if a species is reported at a location outside of its known range.

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 aquatic locations excluding sites located in D01, 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.

4.2 *inv_metabarcodeTaxonList*, *inv_metabarcodeTaxonomy*, and *zoo_metabarcodeTaxonomy*

The sequences presented in the Macroinvertebrate Metabarcoding and Zooplankton Metabarcoding data products are not reliant on the master taxonomy list “MACROINVERTEBRATE” (http://data.neonscience.org/static/taxon.html). Rather, they use taxonomic names and morphospecies identifiers from existing sequence databases identified in the inv_metabarcodeTaxonList table, such as GenBank.

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. The field data entry workflow for collecting zooplankton field data is diagrammed in Figure 4.

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 documents NEON Raw Data Validation for Macroinvertebrate metabarcoding (DP0.20126.001) (AD[03]), 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[12]).

5.2 Automated Data Processing Steps

Following data entry into a mobile application or 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 Sequencing Data

Sequencing data are generated in batches of multiple samples. After sequencing, the multiplexed sequence data are parsed into separate files on a per sample basis. For each sample, minimum quality criteria must be met in order to accept the data for the sample. The general criteria include meeting a minimum sequencing depth (e.g. number of sequences per sample), a maximum number of ambiguous base calls, and a minimum quality score. The actual criteria may change over time as technology evolves and standards change.

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]).

5.4 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 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.

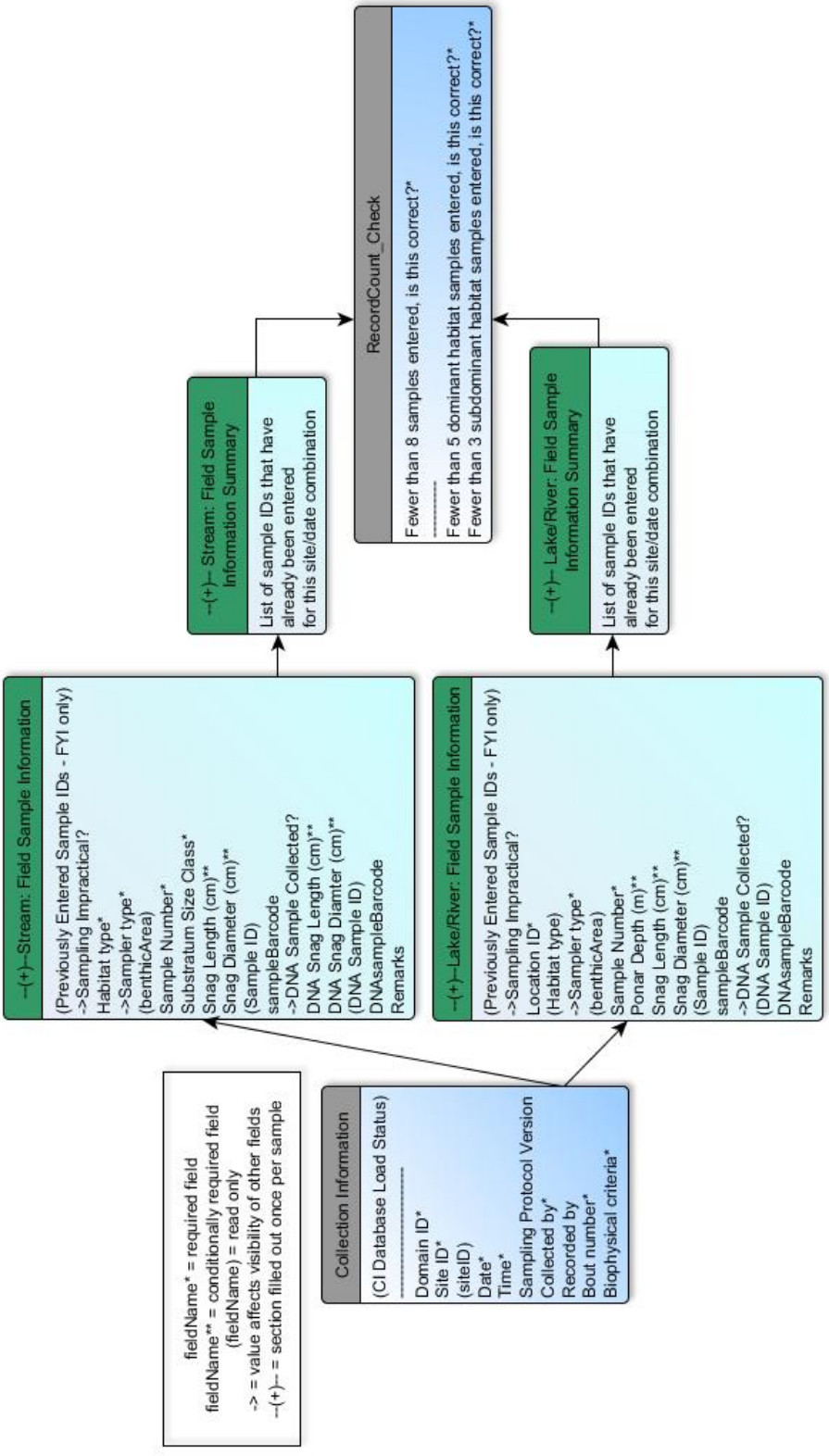

Figure 3: Schematic of the applications used by field technicians to enter macroinvertebrate field data. Boxes with a gray header indicate general information that is filled out one time per sampling event (bout), and applies to all of the subsequent data. The boxes in green indicate data that are populated for each stream transect and point collected.

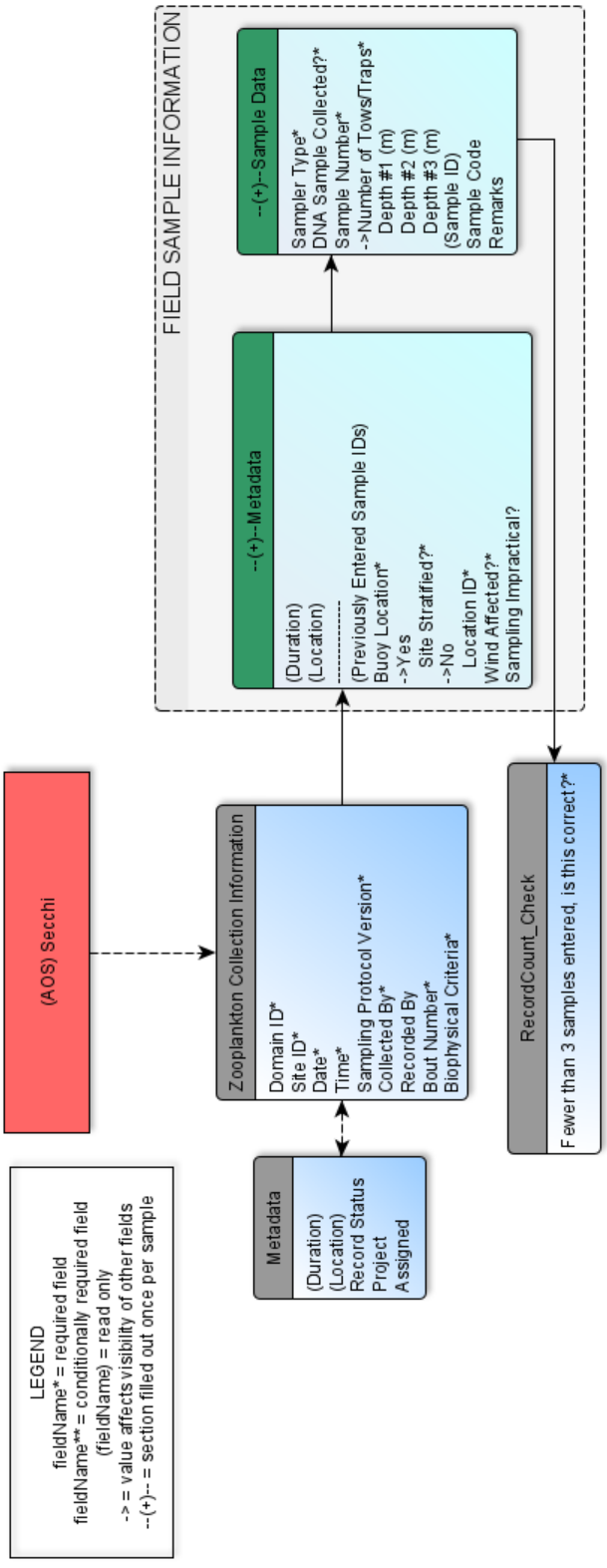

Figure 4: Schematic of the applications used by field technicians to enter zooplankton field data. Boxes with a gray header indicate general information that is filled out one time per sampling event (bout), and applies to all of the subsequent data. The boxes in green indicate data that are populated for each stream transect and point collected.

5.5 Quality Flagging

The **dataQF** field in each data record is a quality flag for known errors applying to the record. Please see the table below for an explanation of **dataQF** codes specific to this product.

Table 1: Descriptions of the dataQF codes for quality flagging

fieldName	value	definition
dataQF	legacyData	Data recorded using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow
dataQF	low extraction efficiency	Extraction method resulted in lower than expected detection of the target taxa, resulting in differences between data from different labs
dataQF	kicknet with incorrect mouth size used	A kicknet with opening 18” x 9” was used rather than the 20” x 13” net required in the field protocol (NEON.DOC.003046), the macroinvertebrate community may be underestimated for these samples

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)

5.6 Analytical Facility Data Quality

Data analyses conducted on sequencing data conform to the current data quality standards used by practitioners. Each metadata table includes a variable, called **qaqcStatus**, in which the laboratory can indicate sample processing issues. Any records with a qaqcStatus = “Fail” should also be accompanied by free-form notes in the “remarks” variable.

Records with the **dataQF** flag “low extraction efficiency” appear to have had extractions performed that optimized bacterial DNA rather than macroinvertebrate and zooplankton DNA, resulting in inefficient detection of the target taxa. See the lab SOPs for more details on extraction and sequencing methods.

6 REFERENCES

Armougom F., and R. Didier. 2009. Exploring microbial diversity using 16S rRNA high-throughput methods. Journal of Computer Science and Systems Biology 2:74-92. https://doi.org/10.4172/jcsb.1000019.

Biggs, B. J. F., R. A. Smith, and M. J. Duncan. 1999. Velocity and sediment disturbance of periphyton in headwater streams: biomass and metabolism. Journal of the North American Benthological Society 18: 222-241.

Brooks, S. S. and A. J. Boulton. 1991. Recolonization dynamics of benthic macroinvertebrates after artificial and natural disturbances in an Australian temporary stream. Australian Journal of Marine and Freshwater Research 42:295-308.

Gibson J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei. 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS One. 10: e0138432.

Hauer, F. R. and V. H. Resh. 2006. Macroinvertebrates. Pages 435-463 in F. R. Hauer and G. A. Lamberti, editors. Methods in Stream Ecology, Second Edition. Academic Press, Boston, MA.

Klindworth A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, and F. O. Glockner. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research 41:e1-e1.

Matthaei, C. D., U. Uhlinger, E. I. Meyer, and A. Frutiger. 1996. Recolonization by benthic invertebrates after experimental disturbance in a Swiss prealpine river. Freshwater Biology 35: 233-248.

Meyer F., D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, et al. 2008. The Metagenomics RAST Server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

Moulton, S. R., II, J. G. Kennen, R. M. Goldstein, and J. A. Hambrook. 2002. Revised protocols for sampling algal, invertebrate, and fish communities as part of the National Water-Quality Assessment Program. Open-File Report 02-150. U.S. Geological Survey, Reston, VA.

USEPA. 2012a. National Lakes Assessment Program, Field Operations Manual.

USEPA. 2012b. Sampling Procedures for the Great Lakes.