AOS/TOS PROTOCOL AND PROCEDURE: DATA MANAGEMENT

PREPARED BY	ORGANIZATION	DATE
Cody Flagg	SCI	04/02/2019
Katherine M. Thibault	SCI	1/05/2015
Sarah C. Elmendorf	DPS	1/05/2015
Courtney Meier	SCI	1/02/2018

APPROVALS	ORGANIZATION	APPROVAL DATE
Kate Thibault	SCI	04/04/2019
Mike Stewart	PSE	04/06/2019

RELEASED BY	ORGANIZATION	RELEASE DATE
Anne Balsley	CM	04/16/2019

See configuration management system for approval history.

The National Ecological Observatory Network is a project solely funded by the National Science Foundation and managed under cooperative agreement by Battelle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Change Record

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
A	05/30/2014	ECO-01835	Initial release
B	01/05/2015	ECO-02633	Added information regarding recent developments of web interfaces and mobile applications for data entry
C	04/04/2016	ECO-03640	• Updated formatting to latest Protocol/SOP template (previous version did not have SOP sections) • Added more text to Background section • Added clarifying steps for data entry QA procedures • Updated acronyms in table 2 • Section 8.2 – updates to best practices section
D	03/09/2018	ECO-05383	• Changed title from “Manual Data Transcription” to “Data Management” to better reflect the contents • Added text in Background section to address mobile data collection with no paper datasheets • Expanded Method section that provides more rationale and language describing data quality concepts • Updated definitions to better match digital data collection concepts • Previous Table 1 was deleted as it was no longer relevant. • Deleted old Table 2, no longer relevant • Added Table 4, Field Sampling Completeness table for Process Quality review • Added Table 5, Protocol Sampling Completeness table for Process Quality review • Replaced Figures 1 with a more up to date version of the data ingest process. • Added Figure 2, workflow diagram from field data collection to data quality review. • Re-arranged SOPs to emphasize the mobile/digital elements of data collection now (SOPs originally started with paper datasheet collection workflow) • Added SOP A: general overview of data collection and data quality review process

REVISION	DATE	ECO #	DESCRIPTION OF CHANGE
			• Added SOP B: data quality methods for detecting duplicate records • Added SOP C: data quality methods for assessing process quality (complete record sets) • Removed old Appendix A "Best Practices for Data Entry and Data Quality Management" • Removed old Appendix B "NEON Specific Guidelines and Tips for Taxonomy related Data Entry" • Old Appendix C, "Data entry training checklist" is now Appendix A • Added Appendix B "QA/QC for Digital Hemispherical Photos" • Added Appendix C "Vegetation Structure QA/QC"
E	04/16/2019	ECO-06020	• Section 3.1.2: Added paragraph about how checklists should be developed and used for QA/QC • Section 3.1.3: Added a new table as "Table 1" – general description of parser validation rules. • Section 3.1.3: Added more descriptive text about the data ingest process. • SOP E: New SOP outlining how to create and use QC checklists, and when they should be implemented. • Removed Appendix B "QA/QC for Digital Hemispherical Photos", this should go into the new protocol specific QC checklists that are being developed. • Removed Appendix C "Vegetation Structure QA/QC", this should go into the new protocol specific QC checklists that are being developed. • Added Appendix B "How to Export Fulcrum Data"

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 OVERVIEW

1.1 Background

This document provides a cross-protocol set of data management instructions for the NEON Terrestrial and Aquatic Observation Systems. It includes guidelines for transcription and storage when field and lab data are recorded with paper datasheets, best practices for data collected on mobile devices, and a framework for data quality control and assurance measures that should be applied before, during, and after data collection.

Data management is the application of a "…consistent methodology that ensures the deployment of timely and trusted data" (Fisher 2009). There are two primary goals that staff should understand before engaging in the data management process. The first goal is to ensure that high quality data are collected, maintained, and delivered to end users by preventing or identifying data quality issues. *Data quality* issues can be defined as any condition that is an obstacle to the data consumer's use of those data (Sebastian-Coleman 2013). Data management in this context is the process by which we prevent, remove, or reduce obstacles to the effective use of data. The second goal of data management is therefore to establish a responsive process that identifies, measures, tracks, and resolves data quality issues.

The aim of quality control (QC) during data management is to prevent the introduction of errors throughout all stages of data collection, transcription (where relevant), and storage. The purpose of quality assurance (QA) is to detect and correct errors, and prevent future data quality issues. Identifying and resolving data quality issues early in data collection processes is particularly important to NEON's strategic goal of providing standardized, long-term data sets to the ecological community. It is critical that personnel understand the data quality process and their role in it at NEON.

1.2 Scope

This document provides a change-controlled version of Observatory protocols and procedures. Documentation of content changes (i.e. changes in particular tasks or safety practices) will occur via this change-controlled document, not through field manuals or training materials.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

Applicable documents contain higher-level information that is implemented in the current document. Examples include designs, plans, or standards.

AD[01]	NEON.DOC.001155	NEON Training Plan
AD[02]	NEON.DOC.014051	Field Audit Plan
AD[03]	NEON.DOC.000824	Data and Data Product Quality Assurance and Control Plan

2.2 Reference Documents

Reference documents contain information that supports or complements the current document. Examples include related protocols, datasheets, or general-information references.

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.000243	NEON Glossary of Terms
RD[03]	NEON.DOC.005003	NEON Scientific Data Products Catalog

2.3 Acronyms

Acronym	Definition
AOS	Aquatic Observation System
DEA	Data entry application
DSF	Domain Support Facility
FOPS	Field Operations
MDR	Mobile data recorder
PDR	Processed data repository
QA	Quality assurance
QC	Quality control
SSL	Sampling Support Library
TOS	Terrestrial Observation System

2.4 Definitions

Common terms used throughout this document are defined here, in alphabetical order.

Child record: in data entry applications, the nested form/screen where sub-sample data or multiple observations are recorded; child records inherit metadata from the parent record. The paper datasheet analog would be the rows where sample, trap, or observation specific data are recorded.

Data entry application (DEA): An electronic, protocol-specific user interface created by NEON Science to provide controlled data entry. Can be accessed through mobile devices or desktop/laptop computers. Will be used by NEON FOPs staff to submit data to the Processed Data Repository (PDR).

Data quality dimension: characteristics of data that can be measured through which *data quality* can be described and quantified.

Data quality context: the description of when and where data quality tasks are conducted within a workflow.

Data quality focus: the combination of data quality dimensions that build towards organizational data quality goals (example: a process quality focus is the assessment of both sampling completeness and timeliness).

Downstream application: an application that *receives* metadata from another application

Field datasheet: A pdf document to be printed and filled out by hand in the field or the lab.

Load Delay: the application/protocol specific number of days before records are ingested and evaluated by the parser. Number of days are counted from date of record creation. Records that successfully "pass" the parser validation rules are set to a read-only state; rejected records remain editable.

Load Group: the set of applications whose data are ingested at the same time. Applications related to the same protocol are generally within the same load group (e.g. Litter: Trap Deployment, Litter: Field Sampling, and Litter: Lab Mass Data).

Metadata: information that describes where observations or samples were collected, when those data were collected, and who collected them.

Mobile data recorder (MDR): handheld, field-portable equipment that runs protocol-specific data entry applications created by NEON Science.

NEON intranet: a generic reference to the NEON Intranet pages that house links to training materials, instructions, protocols, datasheets, and other documents that are shared among FOPs, TOS, and AOS.

Observer: the person that performs a measurement or observation.

Orphan Record: a record whose unique identifier (e.g. sampleID, individualID etc.) is not logically traceable to an upstream record. Field data are often related to downstream domain lab and/or external lab generated data. An example orphan record would be a lab sample whose sampleID does not match metadata for a field collection event (e.g. a Soil Moisture sample "GRSM_006-O-11.5-34-20180409-sm" *should* match an upstream Soil Core Collection sample "GRSM_006-O-11.5-34-20180409"; if it does not, the Soil Moisture record is an orphan.).

Parent record: in data entry applications, the first form/screen where metadata are recorded; parent records often contain and are linked to child records. The paper datasheet analog would be the

"header" section where sampling location, sampling date, and field personnel information is recorded.

Parser: data validation software that evaluates Fulcrum records against protocol specific rules. Either "passes" and locks records that successfully pass validation and moves data on for web publication, or "rejects" records that fail validation from the publication process.

Quality Assurance: methods for detecting and correcting errors, and for preventing future data quality issues.

Quality Control: methods for preventing the introduction of errors throughout all stages of data collection, transcription, and storage.

Record: the collection of data values that describe a sampling entity or event. If data are viewed in a spreadsheet or tabular format, each row is a record and each column is an attribute of that entity or event.

Recorder: the person who enters the data.

Syncing/synchronization: the process of transferring digital records from a DEA to a cloud database for long-term storage. On mobile devices, this specifically refers to the act of pressing the "sync" button. On laptop or desktop computers, syncing occurs as soon as a record is saved.

Upstream application: an application that is the *source* of metadata such as plotID, collectDate, sampleID etc.

3 METHOD

This protocol provides Standard Operating Procedures (SOPs) for all aspects of data management that are relevant to field staff. SOPs A, B, and C describe how data collected with mobile data recorders should be managed and reviewed. SOP D addresses management and review of data collected on paper datasheets. Lastly, instructions found within each protocol, under the "Data Entry and Verification" SOP, may exist and supplement procedures here. SOP E provides instructions for how data quality checklists should be created and implemented.

3.1.1 Data Quality Framework

Data quality can be defined as the degree to which data meet the expectations of data consumers per the intended use of those data (Sebastian-Coleman 2013). More specifically, the proposed framework helps to clarify answering the *what* (data quality dimensions), where & when (data quality context), and why (data quality focus) of the entire data quality process. While these expectations are not objective and can vary widely among the broad range of potential NEON data users, there are dozens of data quality attributes that can be measured, reported, and assessed by users. These attributes can be directly measured by NEON staff as well, and are summarized as **data quality dimensions** (from Sebastian-Coleman 2013):

1. **Completeness**: the expected data exist and are complete according to pre-defined characteristics (e.g. all relevant attributes are measured and the amount of data is correct)
2. **Timeliness**: data are collected, processed, and delivered at the expected time according to a set schedule or when an event occurs
3. **Validity**: data conform to the expected syntax, allowable type (string, integer, date etc.), ranges, or other pre-defined rules
4. **Consistency**: the absence of variance or change in comparison to some pre-defined expectation
5. **Integrity**: data contain all relevant relationship linkages (e.g., no orphans or missing child records)

Where and when data quality is controlled and assessed is just as important to consider as *how* to measure quality, and is referred to as the **data quality context**. At NEON, the following contexts should be familiar to experienced staff:

- **Pre-Collection QA/QC**: activities that take place prior to data collection, aimed at maximizing data collection consistency e.g. protocol training, observer calibration (for observations e.g. phenophase evaluation or estimating percent cover), etc.
- **Collection QA/QC** (or "field QA/QC"): activities that take place during or immediately after data collection, but before leaving a field site, aimed at maximizing data validity and completeness
- **Post-Collection QA/QC** (or "office QA/QC"): activities that generally take place after field collection, and are designed to maintain data completeness, sampling timeliness, and between bout observation consistency

Measuring multiple data quality dimensions simultaneously is important in assessing the quality of underlying processes of data collection within an organization. These are referred to as a **data quality focus**, of which the OS subsystem is primarily interested in or focused on:

- **Referential Integrity (validity and integrity)**: the absence of duplicate, orphaned, and/or 'childless' records from datasets
- **Process Quality (completeness and timeliness):** the presence of all expected data and/or samples within a well-defined time interval
- **Data Plausibility (validity and consistency):** a qualitative assessment of whether data are "reasonable" or not. Data Plausibility issues and methods are NOT addressed in this document, but should be documented within protocols as a "Data QA/QC Checklist" (as indicated in SOP E).

3.1.2 Data Quality Review Workflow

Mobile applications are the preferred mechanism for data collection and entry. The first stage of the process is to review critical field metadata values soon after collection, preferably before leaving a field site. Critical metadata values, such as plotID or collection date, are important data that cannot be easily inferred and/or corrected after staff leave a field site, and include data that pass DEA validation rules but are ***inaccurate*** or ***incorrect*** in some way (see **Box 1** and SOP for specifics).

It is recommended that field staff use digital tools to detect, report, and correct issues if tools are provided. If a tool has not been provided, it is recommended that data be first analyzed with filters and Pivot Tables in Microsoft Excel to summarize patterns in large datasets. Viewing raw data can be useful, but is the least desirable option because it's easy to overlook errors when faced with hundreds of data values.

Referential quality checks are concerned with maintaining the "referential integrity" of datasets by ensuring that ingested records are free of potentially misleading data (e.g. duplicate records with slightly different data values are confusing to end users) and have concrete relations to relevant records (e.g. records have all necessary metadata). Duplicate records are a major concern as even a single duplicate record will cause an entire batch of data to fail upload for many NEON data products. Duplicate records can be identified as the presence of two or more records in a data set with the exact same sample identifier.

Ideally, field staff will review data for referential quality (e.g., duplicates) once an entire bout's worth of data have been entered into (or collected with) a protocol-specific DEA. The exception is for bouts that extend for long periods (i.e. bouts longer than 2 weeks), in which case data reviews should occur before the bout is completed. At a minimum, data should be reviewed before the load delay is reached for a data product, as this simplifies the data editing and record deletion process. Technicians will use the Magpie data viewer to conduct QA/QC tasks. Magpie can be accessed via the Data Management page of the Sampling Support Library (SSL) and contains instructions on how to use the interface. Technicians

will assess the results of each query conducted with Magpie, then determine whether corrective action needs to be taken, such as submitting record deletion requests or correcting data.

Identifying process quality issues allows NEON to improve data collection procedures over time while also ensuring that all expected data/samples are collected and published. Incomplete data sets can be particularly detrimental to scientific end users as fewer samples or data points decrease the ability to detect significant ecological changes or trends. Analyzing process quality includes assessing whether domains and field staff are correctly implementing NEON protocols, and encompasses 'bout-level' metrics such as sampling completeness and sampling timeliness, as prescribed by protocols. Quantitatively, analyzing your data for process quality issues means checking that:

a) Expected number of field records are present per site and protocol ("Field Sampling Completeness")
b) Expected number of records are present for all SOPs in a protocol per sampling location, bout, and protocol ("Protocol Sampling Completeness")

Finally, domain staff should develop or adopt existing QA/QC checklists for the protocols they specialize in. Every sampling protocol should contain a list detailing key data attributes to review before data are ingested. These lists should be found in the "Quick Reference" section of a protocol's appendix. Many checklist items describe important aspects of a data product's quality that are not, and may never be, detected by the OS Parser. For example the Parser cannot inspect the photographic quality of images collected for the measurement of Leaf Area Index, thus the QA/QC list highlights the following checks: "Are images in focus? Is ISO within acceptable range? Is f/stop within acceptable range?" Checklists should be used by staff at all levels throughout the data review process, and are intended to ensure consistency, to convey best practices, to help avoid poor decision making, and to sustain patterns of success.

3.1.3 Data Validation and Ingest Process

Validations

Field staff must understand how data are transferred, or "ingested", from data entry applications to the NEON Web Portal as digital records are only directly editable for a limited time frame. Automated data validation checks are documented in a data product's "Ingest Workbook", which details a two stage process. The first round of data validation occurs directly within Fulcrum applications. Data that do not meet the expectations of these rules are generally prevented from being saved and synchronized with the cloud database. Fulcrum validation rules are limited to checking data within a single parent record, thus rarely check for referential integrity issues. For example, these validation rules cannot check for duplicate sampleID or barcode values across multiple parent records. A piece of software known as the "OS Parser" (or "parser" for short) carries out the second round of validations and makes up for the limitations of the Fulcrum rules. The parser executes several referential integrity checks to prevent duplicates and orphaned data from entering the database in addition to checking the Fulcrum rules

again (Table 1). This second check of Fulcrum rules is done to ensure that the applications have been designed and implemented correctly by application developers.

Timing ("load delays")

Every record entered into a DEA is "stamped" with a creation date (**Figure 1**). A "load delay" countdown begins, starting from the creation date, once these data are synchronized with the cloud database. The load delay defines how long a record can be edited by field staff and varies from 14 to 365 days. Once a record exceeds the load delay, the data are tested by the OS Parser. Records that "pass" Parser validation are then locked and the data eventually appear on the Web Portal. Field staff can view but not directly edit locked records. Records that "fail" validation generate error messages that are collated across protocols and distributed to FOPS and HQ staff by a tool named Canary. The pipeline attempts to re-ingest failed records once a week and will generate the same error messages until data issues are addressed. Failed records must be corrected by field staff. Changes to locked and loaded data records may be necessary if staff determine the need for updates (e.g. a valid taxonID was selected, but is known to be the incorrect sub-species for a particular site) or detect data errors not caught by the parser (e.g. a valid but incorrect collection date was entered). These updates must be submitted through the Magpie application (i.e. "Update Requests") and are handled by staff scientists.

Difficulties

One major caveat of the ingest is that records are not ingested individually; instead, records are ingested in 'batches' or 'sets' referred to as 'load groups'. Load groups contain all the records from a suite of related applications for a given domain. This means that a single record fail results in all related records within and across applications failing at the same time (all records will also receive a load_status of "PARSE_FAIL", which can be misleading for troubleshooting). All related records are rejected as there may be duplicate records present that need to be resolved or deleted before the data can be ingested. OS data are sometimes heavily interrelated across field, domain lab, external lab, and shipping applications. Since metadata are "inherited" across these applications, errors in one application may cause the Parser to flag records in other linked applications. This inter-application complexity can make resolving parser errors difficult if one does not understand what the Parser is checking for or how metadata are related across applications. Finally, post-parser "Update Requests" can be very time consuming and difficult to work through because records can no longer be edited through Fulcrum applications.

Table 1. OS Parser Data Validation Rules

Method	Checks for…	Error Message
Geographic Range Constraint (integrity, validity)	Valid plotID names for a specific protocol are selected. Sampling cannot occur at certain types of plots or transects.	NAMED_LOCATION_TYPE
Sample Identifier Check (integrity)	Whether sample identifiers are unique within the system and/or whether downstream sample identifiers correctly match upstream identifiers e.g. checks for duplicate and orphan records.	"Samples does not exist", "Sample already exists", ReferenceCount
Barcode Check (integrity)	Whether scanned barcodes are unique, whether a barcode has been associated with more than one sample type (e.g. a barcode scanned in ALG and again in CFC would fail validation)	Has different tags, has different barcodes, configuration error changing tag
Taxonomic Value Constraint (validity)	Only animal or plant taxa that have been identified to exist within a domain's geographic boundaries are selected.	ELEMENT_OF
Required Fields (completeness)	Whether required fields have been filled out; conditionally dependent on data entered, whether certain fields have been left blank. Fields are usually required as a way of delivering a minimum usable data product for end users (e.g. many data are unusable if collection dates are missing).	REQUIRE, REQUIRE_NULL
List of Values Constraint (validity)	Whether selected data values match against a pre-specified list. More of a check on the application's design because end users cannot enter custom values.	LOV
Numeric Range Constraint (validity)	Whether entered values fall within a specified range.	_THAN

4 SCHEDULE

4.1 Data Entry and Review

It is preferable to use MDRs and NEON data entry applications to capture data electronically in real-time whenever possible. Data collected on an MDR should be synchronized **as soon as possible.**

Data collected on **paper datasheets should be entered within 14 days of collection** for protocols with long bouts (e.g., Vegetation Structure) *or* within 14 days of the end of a bout for protocols with short bouts (e.g., Mosquitoes).

All digital records, whether originally recorded with an MDR or paper datasheet, must be reviewed and edited ***before*** the load delay occurs. Load delays per protocol are defined on the "About" page of the Magpie application: den-raven-1.ci.neoninternal.org/Fulcrum-QA/

The following tasks should be prioritized from highest to lowest priority, in descending order, **within 14 days of collection or the end of a bout**:

- Field datasheets and data entry application records should be quickly reviewed for critical errors by technicians on a daily basis (SOP A.1, step 2). Critical errors are incorrectly recorded sampling locations and sampling dates that cannot be accurately corrected after leaving the field.
- Sync MDR data when a web/data connection is available (SOP A.1)

The following review tasks should be prioritized from highest to lowest priority, in descending order, and completed within the **load delay period**:

- Perform referential integrity reviews for a given protocol before data are ingested and locked (SOP A)
- Perform a process quality review for a given protocol at the end of a sampling bout and field season (0)
- Perform additional quality control procedures outlined in protocol specific "Data QA/QC" checklists (SOP E)
- Scan and store field datasheets (where applicable) on a weekly or bout-level interval (SOP D.2)
- Manually enter paper data (where applicable), starting with the earliest collection date. *Every paper data entry bout should include paper datasheet data quality checking* (SOP D.4)

5 PERSONNEL

Technicians entering data should familiarize themselves with the protocol and its datasheets for which they are entering data. As a best practice, technicians transcribing paper data sheets should be those who collected the field data, or be personnel trained in the specific protocol. Lead field staff should take a guiding role in directing all data management activities, from paper datasheet data entry and data quality checking (DQC) to data quality review.

5.1 Training Requirements

Field staff should have basic knowledge of the software used to host data entry applications and should have familiarized themselves with the relevant data ingest sheets and protocol-specific applications involved in their work.

5.2 Specialized Skills

Prior experience viewing, sorting, filtering, and manipulating data in spreadsheet software such as Microsoft Excel is desirable but not required.

6 STANDARD OPERATING PROCEDURES

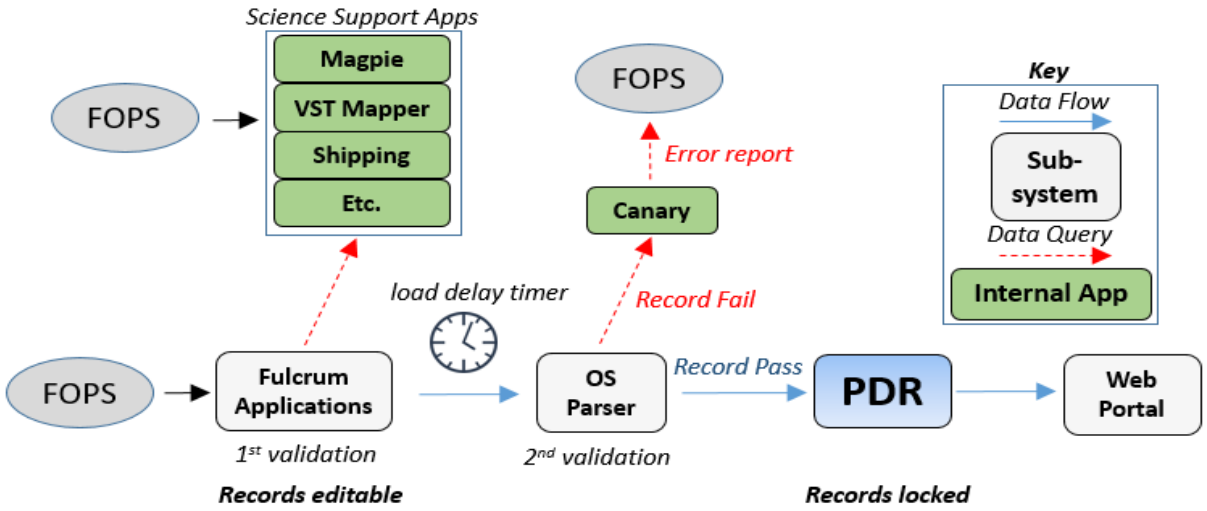

Figure 1. Observatory System data ingest process. Records begin moving through the ingest process once the "load delay", or number of days since record creation, is exceeded. Data are automatically validated twice in the process, before being stored in the Processed Data Repository (PDR) and accessed via the Web Portal. Data for specific protocols are manually validated with various "Science Support Applications". The Canary application analyzes and distributes error messages to Field Science and HQ staff when records fail to pass through the OS Parser.

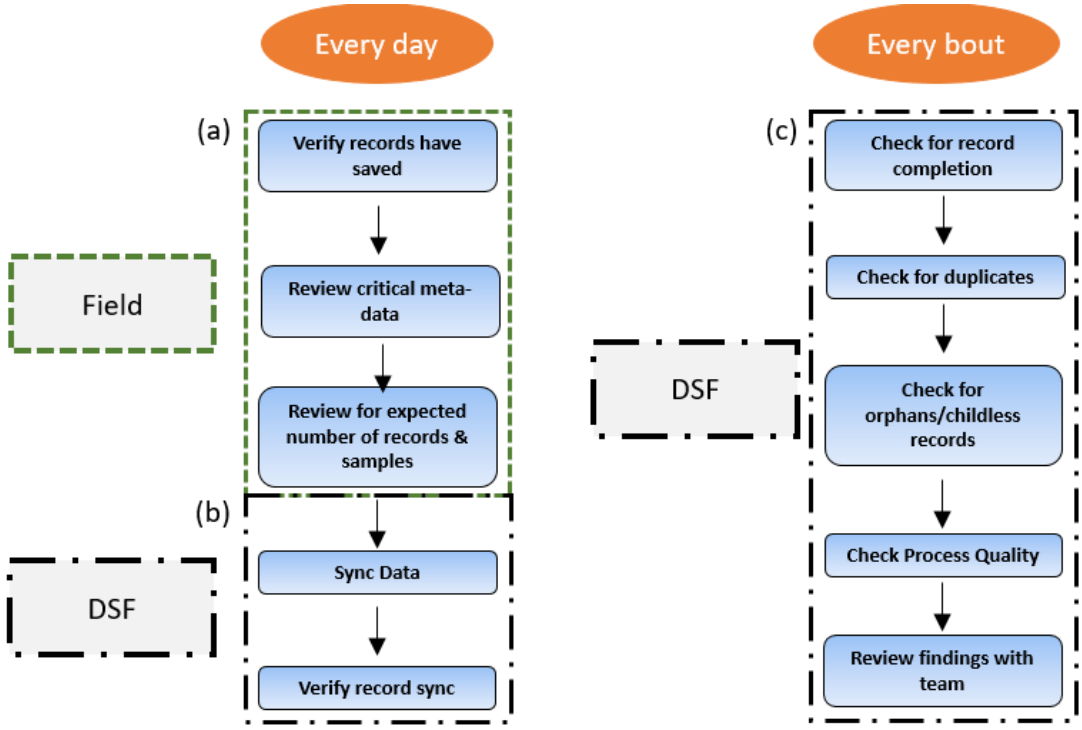

Figure 2. Workflow diagram for reviewing data quality per day and per bout. Dashed boxes represent *where* data should be reviewed, orange circles represent *when* in the workflow review steps should happen.

SOP A Mobile Data Recorder: General Data Quality Workflow for all Protocols

Ensure that the person conducting data entry has read through the training materials associated with the protocol-specific data entry application, and that they have practiced entering data using the *Training* version of that application – i.e., the [CERT] version. A training checklist for mobile data recorder data entry is provided in **Appendix A**. Staff entering data for the first time should work through the checklist, in combination with reading the SOPs below, to ensure an understanding of the NEON OS Data Quality Workflow.

A.1 Every Day: Before Leaving a Field Site

Context: Collection QA/QC

The component of the Quality Workflow that is implemented daily before leaving the field site is shown in **Figure 2a**. Steps include:

1. **Verify that entered data have been saved on the MDR**
    a. For parent-child records, be sure to save the **parent** record. Saving child records does NOT mean that a parent record has been saved.
        i. ***NOTE***: It is a common misunderstanding that the data entry application will automatically save any entered data. A person who is not the data recorder should double-check that the recorder has properly saved all data before leaving a field site. All too often, data are entered and then discarded by accident only for the data loss to be recognized after leaving a field site.
        ii. ***WARNING***: Records can be saved in a temporary state called a "draft" when one or more required fields are left blank (appears as a yellow pencil icon on mobile devices). ***Draft records are never synced to the cloud*** and should be completed as soon as possible to prevent data loss – i.e., the draft record will only exist on the device used to collect data.
2. **Review critical metadata fields for accuracy and completeness**. This step is important to prevent incorrect, but technically valid data from being entered (Box 1)
    a. Check for valid and correct *sampling location* values (e.g. siteID, plotID, clipCellNumber, trapID etc.)
    b. Check for valid *sampling date and time* values (e.g. setDate, collectDate, setTime, collectTime etc.)
    c. Check for valid *event identifier* values (e.g. boutNumber, yearBoutBegan, eventID etc.)
    d. Check that the expected number of records has been created and saved on the MDR
        i. ***Example***:
            1. For many data entry applications, one parent record is often created per sampling location.

a. If the field crew is expected to sample 10 mosquito traps in one day, there should be 10 mosquito records on the MDR

2. In other data entry applications, one parent record represents a single sampling location and the child records represent sub-samples within that same location.
    a. If a field crew is visiting three plots to take digital hemispherical photos there should be:
        i. Three parent records (one per plot)
        ii. Either 12 or 24 child records per parent record (12 or 24 photos taken per plot)

Box 1. Inaccurate but technically valid data

Definition: Entered data values that pass DEA validation rules but are mis-recorded or factually incorrect. Inaccurate data records are particularly troublesome because while they can often be identified as incorrect, how to correct the data is not always obvious or possible. Delaying the data review process leads to the loss of useful information from field staff that might help rectify errors.

- ***Example***: Data recorded for plot "X" when observations are actually from plot "Y"
    - Actual plotID = "CPER_012"; recorded plotID = "CPER_014"
    - *Results in*…a data set with fewer records than expected AND duplicate records; lower process quality and spatio-temporal data gaps that reduce statistical power
- ***Example***: A recorded tree diameter that is within range but an order of magnitude off from the actual measurement
    - Actual Tree Diameter = "12 cm"; Recorded Tree Diameter = "120 cm" (min value = 1 cm, max value = 400 cm)
    - *Results in*…a data set with more outliers; reduced confidence in the plausibility of the data

3. **If field samples were collected**
    a. Check that the number of samples matches the number of recorded digital records
    b. Check a subset of sample labels and/or barcodes for accuracy
    c. At a minimum, check 10% of the total samples or 10 samples (whichever is greater)
        i. For samples with barcodes, the barcodes can be scanned with the MDR via the "Search" function or using a handheld barcode scanner
            1. Records with a matching barcode will appear in the record list *if they exist*
                a. **If no records** are returned after scanning a barcode, this means the barcode wasn't initally scanned correctly or at all, and you should find the appropriate record and ensure that the barcode is scanned with that record
                b. **For records that appear** after scanning, verify that the record's metadata match the information on the sample
                c. **If more than one record appears** after scanning, a barcode has likely been scanned more than once and associated with different unique samples. This is considered a duplicate record situation, and should be resolved before data are loaded
        ii. For non-barcoded samples, the information on the sample should match metadata in the record

A.2 Every Day: Returning to the Domain Support Facility or Field House

Context: Post-Collection QA/QC

The component of the Quality Workflow that is implemented daily upon returning to the DSF or Field House is shown in **Figure 2b**. Steps include:

1. **Sync the MDR's data**
    a. Be sure that the device is connected to a secure Wi-Fi network or has a data connection before attempting to sync
    b. The synchronization mechanism can fail for multiple reasons, meaning that data are not fully transferred from the MDR to the cloud database
        i. ***NOTE:*** Un-synced draft records are isolated to the device they were collected on. They will not appear on another device or in a web browser because the data have not been properly transferred.
2. **Verify that all MDR data have been synced**
    a. Use a web browser to check that records have transferred from the MDR to the cloud
        i. Navigate to the appropriate DEA (application names are identical between the MDR and web browser)
        ii. Sort the data table by the "Created date" column (this is the date and time a record was created) by selecting "Sort Dec -> Jan"
        iii. The most recently created records will appear at the top of the table
        iv. Locate and identify the records that should have been synced from the MDR
            1. If not all expected records appear in the table, double-check the MDR to see if the syncing process failed and try to sync the device again
3. **If physical specimens were collected**
    a. Transfer samples to the appropriate storage containers and locations
    b. Record this information in a sample inventory
    c. Verify that the number of samples stored matches the expected number of records

A.3 Every Bout: Data Quality Assurance

Context: Post-Collection QA/QC

This section gives a brief overview of the rest of the data quality review process, the components of which are shown in **Figure 2c**. Expanded step-by-step details are provided below in 1.1.1.1.1SOP A and 0.

1. **Check that all entered records have been completed**
    a. All data values entered into each record
        i. Certain applications allow users to save records without entering key metadata in order to match the field workflow. These data values should always be recorded even if they are not specifically required in the application, else records will be rejected by the ingest system.
            1. **Example**: Mosquito Collection ("MOS") application records can be saved without a collectDate or collectTime, however records ingested without these values will be rejected by the OS Parser
    b. For protocols with multiple SOPs and data entry applications
        i. Check that all related records have been entered
            1. **Example**: Herbaceous Clip Harvest ("HBP") has two applications: Field and Lab. One Lab record should exist for each Field record where targetTaxaPresent = "Yes".
    c. Be sure to follow additional protocol-specific instructions from the "Data Entry and Verification" SOP within each protocol.
2. **Check for duplicate records**
    a. Use the Magpie application to find and resolve duplicate records (SOP B.1)
3. **Prevent orphan records from being created**
    a. The parser will detect and reject orphan records
    b. There are currently no tools to find orphan records, however field staff can implement procedures that prevent the creation of orphan records (SOP B.2)
4. **Perform quantitative process quality review**
    a. Check that all expected digital records for a given protocol and bout exist
    b. Check that all expected sampling locations for a given protocol and bout have been visited
5. **Perform additional quality control review outlined in protocol specific "Data QA/QC" checklists (SOP E)**
    a. Check dataset attributes that are not detected by the parser

SOP B Referential Integrity Data Quality Review

Context: Post-Collection QA/QC

B.1 Analyze Data for Duplicate Records

Sample identifier fields are commonly named "sampleID", but also include a number of variations that all end with the suffix "ID" as described in the relevant protocols e.g. individualID, subsampleID, fieldSampleID, massSampleID, moistureSampleID, archiveID etc. Sample identifiers are composed of human-readable unique ***metadata*** from each record, such as:

- plotID
- date
- sub-location, and
- sometimes a suffix or prefix related to the module (e.g. "***hbp***.CPER057163").

To analyze data for duplicate records:

1. Navigate to the NEON Magpie application in a web browser (or copy-paste: den-raven-1.ci.neoninternal.org/Fulcrum-QA)
   a) You must be connected to the internal NEON network to access this location
2. Enter the appropriate search parameters for the data set you are checking e.g. **subsystem**, **domainID**, **siteID**, **application**, and **SOP** (if applicable)
   a) For **Query Type**, select "Duplicates by Sample Identifier"
   b) For **Report Type**, select "Deletion Request"
   c) If there are multiple sample identifiers (e.g. sampleID and subsampleID), an additional field will appear for you to indicate which value you'd like to check against
3. *If there are any duplicates*, the application will return a set of records on the screen
   a) Sort the records by the sample identifier field so that duplicate records are ordered together
   b) You must now determine which record is the erroneous record that should be ***fixed (i.e. edited)*** or ***discarded (i.e. reported for deletion)***
      i. ***Records can be discarded or reported for deletion:***
         (1) If every single column value is identical across the duplicate record set (except fulcrum_id or created_at date), you can simply report all but one record for deletion
            (a) If the duplicate record in question is a **child**, you can simply discard it yourself without using Magpie to make a request
               (i) Navigate to the record in Fulcrum
                  1. In a web browser, click on the "X" symbol to the left of the child record's title
               (ii) On a tablet, long pressing on a child record will bring up a "context menu"
                  1. Select the trash bin symbol
                  2. Select "yes" when prompted to delete the record

a. **NOTE**: this can be reversed if you discard the changes to the record before saving

(b) If the duplicate record in question is a **parent**

(i) Select the duplicate record(s) in Magpie

(ii) Select "Report Fulcrum Records for Deletion"

(iii) Headquarters staff will process the deletion request

ii. ***Records can be fixed:***

(1) If not all column values are identical

(2) If this is the case, you must then determine whether the recorder accidentally selected or entered a metadata value that created the duplicate sample identifier (i.e. plotID, collectDate, coordinates, etc.)

(a) For example, the wrong (but valid) plotID was selected for a record on the same date as another duplicate record with the same plotID value

(b) Or, an incorrect but technically valid date was selected for a record from the same plotID

(3) Edit the erroneous records so that the sample identifier is not a duplicate value of another record's sample identifier

B.2 Prevent Creation of Orphan Records

For protocols with linked applications, where metadata are referenced across applications (e.g. a lab record references a field sampling event), changes or corrections to metadata that make up the unique identifier fields in upstream applications can result in "orphaned" records. Orphaned records are a data quality issue because data in "downstream" applications cannot be accurately traced to an "upstream" record. Changes to sampling locations (e.g. plotID, clipCellNumber, soil coordinates) and/or sampling dates (e.g. setDate, collectDate) are the most common source of orphan records, as these values are almost always used to construct a unique sample identifier and data edits are never automatically propagated across applications.

Upstream applications are usually the first place where important metadata are recorded by field staff (plotID, subplotID, sampling date, soil coordinate, clip cell number, sample type etc.). Many protocols generate a unique identifier, typically named "sampleID", that is used to link data across multiple SOPs or data tables. Identifiers that are not **exact matches** across applications result in "orphan" and "childless" records. *Downstream* applications "inherit" metadata from upstream applications, most importantly the sample identifier. Data that were already entered in an upstream application are generally un-editable. This is to prevent users from entering the same metadata values multiple times in different locations, which results in difficult to resolve transcription errors.

1. Have the application user manual on hand (Fulcrum Manuals are currently linked via each protocol page on the SSL).
2. **For each edit to a sampling location or sampling date** in an upstream record (**Table 2**), write down the original sample identifier

a. Using the original sample identifier, locate the downstream record
b. Open the downstream record and update the metadata following instructions from the application's user manual

Table 2. List of protocols that have linked applications. Applications to the left of a table cell are "upstream"; applications to the right are "downstream". Records in downstream applications can have orphan records if corresponding sample information in upstream records is altered. *(p-c) = application has a parent-child structure, child records inherit metadata from the parent and each child must be updated if sample identifier information in the parent is changed.

Upstream ←----------→ **Downstream**

Protocol	App 1	App 2	App 3	App 4
Litter	Trap Deployment	Field Sampling	Lab Mass (p-c)	BGC Subsampling
Vegetation Structure	Plot Metadata	Mapping and Tagging	Apparent Individuals (p-c)	
Vegetation Structure	Plot Metadata	Shrub Groups		
Soils	Soil Core Collection (p-c)	Soil pH		
Soils	Soil Core Collection (p-c)	Soil Moisture		
Soils	Soil Core Collection (p-c)	Metagenomic Pooling		
Soils	Soil Core Collection (p-c)	BGC Subsampling		
Nitrogen Trans.	Soil Core Collection (t-initial, p-c)	Soil Core Collection (t-final, p-c)	Nitrogen Transformation	
Herbaceous Clip	Field Sampling	Lab Masses (p-c)		
Belowground Biomass Coring	Field (p-c)	Weighing	Grinding and Pooling	
Belowground Biomass Coring	Field (p-c)	Dilution		
Ground Beetles	Setting	Collection (p-c)	Sorting (p-c)	Pinning (p-c)
Ground Beetles	Setting	Collection (p-c)	Sorting (p-c)	Archiving and Pooling
Phenology	Field Setup (p-c)	Phenophase Observations		
Phenology	Field Setup (p-c)	Annual Measurements		
Canopy Foliar	Field (p-c)	LMA		
Canopy Foliar	Field (p-c)	Chemistry Subsampling		
AOS Algae	Field	Lab		
AOS Plants	Field	Lab		

SOP C Detecting Process Quality Issues

Context: *Collection QA/QC* (for protocols with bouts that are longer than 2 weeks) or *Post-Collection QA/QC* (***before*** the next bout begins)

C.1 Analyze Process Quality: Field Sampling Completeness

The goal of checking data for "Field Sampling Completeness" is to ensure that all data have been collected from the field *and* are present in the cloud database. Refer to **Box 2** for a more detailed walk-through.

1. From the Fulcrum application dashboard on a web browser:
   a) Click "View Data" for the application you are checking; you should be taken to a table-like view of the application's data
      i. If a table does not appear, click "Table Mode" in the top-right corner of the screen (next to the "+" add record button)
2. If "Clear All Filters" appears in blue text at the top of the table, click the "X" to remove filtering of rows
3. Most protocols have an "*event*" identifier (e.g. boutNumber, boutType, eventID, yearBoutBegan etc.)
   a) If it is not visible, click on the "Column Setup" (three vertical lines) button and search for an event field; check it to make the column appears
   b) For protocols with no event identifier, filter by collectDate, yearBoutBegan, etc.
4. Filter the visible rows to the sampling bout/event you are reviewing
   a) Click the down arrow of the event column, a window appears
   b) Click "Select Specific Values", search for and select an event identifier
   c) Be sure to filter data by a single siteID, as this simplifies record counts and will match the information in **Table 4 (companion spreadsheet, tab 1)**
5. Only records with the specified event identifier will now appear in the web browser table
6. For a given bout, if **the application does not have child or grandchild records**,
   a) The total number of records will appear at the top-left above the table
      i. **Compare this number to what is expected in Table 4, columns E-H,** based on information in the relevant protocol and SOPs listed
         (1) **Columns E** contains rules describing how many digital records should appear per sampling location
            (a) Columns F and G only need to be referenced if an application has nested child and/or grandchild records (see below)
         (2) **Column H is the total expected number of parent records that should appear in the web browser**
      ii. **Evaluate the record count**

(1) If your counts are **BELOW** the expectation, you may be missing records from the bout, missing physical samples from the field, or have incorrectly entered metadata that goes into the event identifier (e.g. yearBoutBegan, siteID, soil coordinate etc.)
   (a) ***NOTE***: You may be able to discover if records have incorrect metadata by sorting records based on the "created date". Records collected in sequence and/or in the same time interval should be adjacent when sorted this way.
(2) If your counts are **ABOVE** the expectation, there may be duplicate records in the dataset and/or too many samples were collected in the field

7. For a given bout, if **the application has child or grandchild records**,
   a) The total number of **parent records** will appear at the top-left above the table
      i. This should equal the number in Column H (Table 4)
   b) For each parent record, open the record and view the number of child and grandchild records present
      i. View the number of child records (Table 4, Column F)
      ii. If present, tally the number of grandchild records per child record (Table 4, Column G)
8. **Evaluate the record counts**
   a) If your counts are **BELOW** the expectation, you may be missing records from the bout and/or missing physical samples from the field
   b) If your counts are **ABOVE** the expectation, there may be duplicate records in the dataset and/or too many samples were collected in the field

C.2 Analyze Process Quality: Protocol Sampling Completeness

The goal of checking for "Protocol Sampling Completeness" is to ensure that all post-field sampling data have been collected, entered, and align with field sampling data. These checks only need to occur if protocols require additional data collection generated from lab procedures. Refer to **Box 2** for a more detailed example.

1. From the Fulcrum application dashboard on a web browser:
   a) Click "View Data" for the first application you are checking; you should be taken to a table-like view of the application's data
      i. If a table does not appear, click "Table Mode" in the top-right corner of the screen (next to the "+" add record button)
      ii. **NOTE**: *This tab will be referred to as the "**upstream**" application (e.g. Field Sampling)*
2. Open another tab in the web browser and location the second application you will be checking
      i. **NOTE**: *This tab will be referred to as the "**downstream**" application (e.g. Lab Mass)*
3. Select "Clear All Filters" for both open tabs
4. Filter the visible rows to the sampling bout/event you are reviewing ***in both tabs***
   a) Click the down arrow of the event column, a window appears
   b) Click "Select Specific Values", search for and select an event identifier

c) Be sure to filter data by a single siteID, as this simplifies record counts and will match the information in **Table 4**

5. For a given bout,
   a) Note the number of upstream records there are after filtering
   b) Note the number of downstream records there are after filtering
6. Evaluate the record counts,
   a) **Table 5 (companion spreadsheet, tab 2)** describes rules for how to count downstream records
      i. These are rules rather than specific numeric values because not all sampling events result in a downstream record, but are instead based on various conditions defined in a protocol (e.g. Lab Mass records are not created if no field sample was collected)
      ii. Applications and protocols not listed in Table 5 do not need to have Process Quality checks performed
   b) In most cases, **the number of parent records in the downstream application should equal the number of parent records in the upstream application**
      i. If your counts are **BELOW** the expectation, you may be missing records in the downstream app that should be present
         (1) Potential reasons for this include:
            (a) Data have been collected, but a device was not fully synced
            (b) Data have been collected, but were recorded on paper and not entered
            (c) Critical upstream metadata may have been changed, causing a mismatch (e.g. siteID or an event identifier were changed in some way)
            (d) Data have not been collected, recorded, or entered
      ii. If your counts are **ABOVE** the expectation, there may be duplicate records present in the upstream or downstream app, or event identifiers may be incorrect (e.g. boutNumber was not used correctly)
      iii. Workflows that deviate from these general guidelines are listed below. See Table 5 for more details.
         (1) Canopy Foliar Sampling (CFC)
         (2) Soil Biogeochemistry and Microbe Sampling (SLS)

Box 2. Detailed example of conducting all Process Quality checks with Litterfall data (Field and Lab data).

1. Open the "**Litter: Field Sampling**" data table in web browser
2. Open the "**Litter: Lab Mass**" data table in another web browser window
3. To assess ***Field Sampling Completeness*** (SOP C.1)
   a. **Organize data**:
      i. In "*Column Setup*" the first four columns are re-arranged as siteID, boutNumber, setDate, trapID
      ii. The rows are sorted by trapID so that duplicates can be spotted
   b. **Filter data:**
      i. In the **siteID** column: select one site, "**UNDE**"
      ii. The event identifier is "**boutNumber**", select a single value "**4**" in boutNumber column
   c. **Assess results:**
      i. UNDE is a forested site that uses 1600 m^2 plots, **Table 4** of the Data Management Protocol suggests that there should be two parent records per plotID and 40 records total (two traps per plot in two random subplots)
      ii. The web browser shows 40 records

SOP D Paper Datasheet Quality Checking

Paper datasheets can be used as a backup data recording method in cases of adverse field conditions, equipment failure, or missing equipment. Data recorded on paper should be entered within 14 days of collection.

D.1 Every Day: Review Field Datasheets before Leaving the Field

1. Review the field datasheets for completeness
   a) Check BOTH sides of the datasheets for extra taxa or notes
   b) Be sure to clarify and annotate any shorthand notes on the datasheet; don't assume the shorthand will be meaningful to staff in the future
   c) Finally, check that all required fields/columns have been filled out
2. Protocol leads should review critical metadata values on datasheets at the end of the sampling day similarly to steps 1 and 2 outlined in SOP A.2:
   a) Review critical metadata fields for accuracy and completeness (see **Box 1**)
      i. Check for valid and correct*sampling location* values (e.g. siteID, plotID, clipCellNumber, trapID etc.)
      ii. Check for valid *sampling date and time* values (e.g. setDate, collectDate, setTime, collectTime etc.)
      iii. Check for valid *event identifier* values (e.g. boutNumber, yearBoutBegan, eventID etc.)
      iv. Check that the expected number of rows has been recorded
3. If physical specimens were collected:
   a) Check that the number of samples matches the number of rows on the datasheet
   b) Check all sample labels for accuracy

D.2 Every Bout: Scan Field Datasheets

1. Scan reviewed and annotated datasheets.
2. Save scanned datasheets in the folder designated on the NEON shared drive using the following naming convention: "moduleAbbreviation_fds_SAMPLINGLOCATION_YYYYMMDDx", where:
   i. 'fds' refers to field datasheet; lds should be used for lab datasheets
   ii. 'YYYYMMDD' is the most recent date on the datasheet;
   iii. 'x' is an optional character to be used in the event of multiple files (representing a-z). Multiple files can result from multiple datasheets needed to accommodate the quantity of data collected on a given date for a given module and/or a module requiring multiple distinct field (or lab) datasheets. If multiple files are not needed, there is no need to use x.
      (1) plotID (i.e. siteID_plotNumber e.g. "HARV_003") can also be appended to the filename, after the siteID, to avoid confusion if multiple field crews have been working at the same site and module for a given date.

Example: phe_fds_CPER_20130710b (this example is 22 characters long, which is less than the maximum length allowed by some models of scanners).

D.3 Every Bout: Data Entry Procedures

1. Be sure you are connected to the NEON network (use of VPN may be necessary at remote sites). Use only supported web browsers, Google Chrome, Mozilla Firefox, or Microsoft EDGE, for data entry. Using an unsupported browser may result in lost data. Supported browsers may change in the future.
2. Login to the webpage that hosts NEON data entry applications (e.g. Fulcrum) and navigate to the appropriate, protocol-specific application. Links and application names can be found on the NEON Intranet.
3. The person conducting data entry must have practiced entering data using the *Training* version of that application.
   a. **Training Applications (CERT)**
      i) The NEON data entry webpage provides a copy of each data entry application for the sole purpose of training. These are labeled with the DEA name plus CERT for certification. Technicians should use these to enter test data for training purposes, as the data are not stored in the NEON data repository.
      ii) New technicians must practice data entry in CERT applications *for each protocol where they plan to enter data*. Domains should consider generating test data that highlight common local data entry issues to assist with training.
      iii) Entered training data should be exported as a .CSV and reviewed by experienced field staff.
4. Once training requirements are met for that application/protocol, data can be entered into the protocol-specific *Production* (PROD) version of the data entry application.
   a. **Production Applications (PROD)**
      i) The data entry webpage provides data entry applications for actual field and lab data entry. These data will be stored in the NEON data repository and published on the data portal.
      ii) It is difficult to locate and fix incorrectly entered data once they have been locked and transitioned to the NEON PDR. *Extra time and effort should thus be devoted to careful data entry and data review*.
5. Data entry applications are customized for each individual protocol. Detailed training for each protocol interface is available in the 'Supporting Documents' section of the protocol page in the SSL, but here are some general rules:
   a. Sync MDR devices often in order to ensure that up-to-date versions of all DEAs are available.

b. Remember that most of the applications have a 'nested' structure, meaning that some of the higher-level information (plotID) is entered once but applies to a group of samples.

c. Data entry through a data entry application is a multi-step process:

i) The '**Create**' button (**+**) will allow you to create a new record.

ii) Once all required information about that record has been entered, use the '**Save**' button (✓) to save the data and store it locally. Note that once records are created, saved, and synced, they cannot be deleted without submitting a request to NEON HQ.

iii) The data entry application will allow you to review and edit data you have recorded until it transitions to the NEON data repository. Review can be accomplished by clicking on the 'Quick View' option for a specific record, while editing can be accomplished using 'Edit.'

iv) Use the **Status** column to identify records that have been flagged as needing further attention – they will be highlighted in red, orange, or other colors.

6. All bouts of data entry must be accompanied by the Paper Datasheet Data Quality Checking procedure, outlined below. <u>This procedure must be conducted before the data entry bout is considered complete.</u>

D.4 Every Bout: Paper Datasheet Quality Checking

1. Field technicians should quality check a minimum of 10% of the entered data records or 10 records, whichever is higher. These targets are either on a per sampling/data entry bout or per SOP level (clarified below). Data quality control is accomplished by comparing entered values to the original paper datasheet values.

a. **Per sampling or data entry bout**: data quality checks should be carried out for each field sampling/data entry bout during a field season. This is done so that errors in data entry and problematic data entry habits can be detected and corrected.

b. **Per SOP**: the number of records reviewed should be considered per protocol SOP rather than all protocol records totaled together.

i. *Example:* After a sampling/data entry bout, a domain has 40 records from Litterfall SOP C (Field Sampling) and 390 records from Litterfall SOP D (Lab Processing – Dry Mass) for bout 3 of the season. Staff should therefore review 10 records from SOP C (because 10% of 40 is only 4 records) and 39 records from SOP D (because 10% of 390 is 39).

2. After a dataset is transcribed, it can be checked in one of two ways – either by exporting as a spreadsheet directly from the DEA, or using the NEON custom data viewer "Magpie". Both options will provide relatively fast ways to compare entered data to paper datasheet values.

3. If choosing to export data as a spreadsheet from the DEA, refer to Appendix B.
4. If using the Magpie NEON custom data viewer, use the instructions provided in the 'About this application' tab to select and filter the records relevant to the bout. Once selected and filtered, the records can be exported in spreadsheet form using the 'Download Query' button. From there, they can be further filtered as described above to facilitate data quality review.
5. Using the filtered data, compare the entered data values of each field to the original paper datasheet with one of the following procedures:
    a. **Read Aloud Review**: Data checking conducted by two technicians: one technician should read the values on the original paper sheet aloud, while the other visually checks against the electronically entered values.
    b. **Independent Data Review**: If it is not possible for two technicians to conduct data quality review together, then a technician who did not enter the current dataset should visually review the dataset.
    c. **Redundant Data Review:** If a separate technician cannot conduct data quality review, then the technician who entered the data should also visually re-check the data.
        i. It is strongly recommended that technicians employ the Read Aloud Review or Independent Data Review procedures over Redundant Data Review.
6. For any records that have errors, navigate to the relevant data entry application and correct them.

SOP E Using Checklists for Quality Control

Incorporating checklists into routine procedures has been shown to reduce the number of serious errors committed in both the aviation and healthcare industries (Clay-Williams and Colligan 2015). Checklists are best suited to situations where performance requires standardization (i.e. reviewing data quality), time is not critical, tasks may be forgotten or skipped, and the number of tasks may be too great to accurately memorize. Checklists are not just a memory aide but should also be considered a tool for discussing important aspects of a procedure with other staff members.

Recent research has suggested that checklists are more likely to be effective and adopted into workflows when teams are actively engaged in the process of developing and adapting them to their particular contexts or work environments (Gillespie and Marshall 2015). Protocol authors therefore have not specified uniformly formatted, "one size fits all" data quality checklists for each protocol that cannot be deviated from. Rather, authors must provide (1) tasks that define the minimum standard for delivering high quality data and (2) effective guidance that supports consistent decision making when minimum data quality standards are not met. In general, minimum data review standards should clearly define the frequency of data review, when data review must occur during a bout, result in binary outcomes (e.g. pass/fail, yes/no), and provide guidance when expected outcomes are not met. Each field office should use the minimum data review standards and decision making guidelines as a starting point for implementing checklists, while also acknowledging and incorporating data quality conditions unique to a domain (e.g. biological soil crust cover estimates are only carried out in D14 and D15).

E.1 Checklist Preparation

1. Consult a protocol's Quick Reference section for key items that have been identified as critical to check.
    a. If a protocol doesn't have a "Data QA/QC Checklist" section yet, the checklist should be located on ServiceNow's Knowledge Base and linked from the Sampling Support Library (SSL).
    b. The minimum data standards specified by the document should be effected for each bout unless otherwise specified.
    c. Any additional checks NOT specified in the "Data QA/QC Checklist" are entirely optional.
2. Two documents will be provided for each protocol's data quality review process:
    a. A Word document that lists out specific details on how to conduct data review, referred to as the "**checklist**".
        i. The checklist provides step-by-step instructions on how to review data, and lists out only the key procedures that need to be carried out
        ii. Anything not listed on the checklist is not considered to be critical to the data product
    b. An Excel spreadsheet that contains abbreviated checklist items referred to as the "worksheet".
        i. The QC worksheet is intended for domain staff to track and communicate where data are in the review process, when review occurred, and who carried out review.

ii. QC worksheet documents *can be modified* by Field Staff as needed, but ***must still contain the QC items listed in the Word document checklist***

3. Review the various checklist worksheet templates provided and adapt it as appropriate for your domain's workflow.
   a. There is not a single right or wrong way to format a checklist
   b. These templates are examples of the different ways a checklist can be implemented
4. Review the Knowledge Base articles that describe the OS Parser rules per protocol
   a. Remember, the Parser will always automatically check data specified in the rules
5. Create a checklist for your domain by combining the items listed in a protocol's Quick Reference section (or Knowledge Base article) with a template
6. Review the content of your checklist worksheets and implementation strategy with fellow domain staff
7. Store the digital version of the checklist in a well-known, easy to find location
8. Incorporate checklists into training

7 REFERENCES

Chapman, A.D. 2005. Principles of Data Quality – Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF.

Fisher, C., Eitel, L., Chengalur-Smith, S., and R. Wang. *Introduction to information quality*. Cambridge, MA: MIT Information Quality Program.

Michener, W.K. and J.W. Brunt. 2000. Data Management Principles, Implementation and Administration. *Ecological Data – Design, Management, and Processing (ed.* W.K. Michener and J.W. Brunt*),* Blackwell Science Ltd, Oxford.

N.J. Van Lanen, C.M. White, J.A. Fogg, and M. F. McLaren. 2012. Integrated Monitoring In Bird Conservation Regions (IMBCR): Data Entry Protocol. Unpublished report. Rocky Mountain Bird Observatory, Brighton, CO, USA. URL: rmbo.org/v3/Portals/0/Documents/Science/Protocols/

Sebastian-Coleman, L. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. 2013. Elsevier Inc: Waltham, MA.

APPENDIX A DATA ENTRY TRAINING CHECKLIST

So, you're ready to begin entering _____ data.
(protocol)

Make sure you complete the following checklist before beginning. A permanent staff member or designated lead technician will need to sign off on your checklist before actual data entry can begin.

- ☐ Read through the Data Management protocol
- ☐ Review Data Management Training Materials
  - ☐ Powerpoint
  - ☐ FAQs/ Common Mistake forms (if available)
- ☐ Read general Data Entry Application (Fulcrum) User Training Guide
- ☐ Read relevant protocol-specific User Guide(s)
- ☐ Enter at least one datasheet, plot or subplot of data into the training data entry application (CERT) for the relevant protocol
- ☐ Talk through the process of:
  - ☐ Annotating datasheets
  - ☐ NEON data entry webpage login, ability to navigate to protocol-specific apps
  - ☐ Appropriate remarks, use of commas and special characters
  - ☐ Manual data transcription DQC process
  - ☐ Where entered and checked datasheets get stored
  - ☐ How to report DQC bouts
  - ☐ Importance of catching and reporting duplicates
  - ☐ Use of the Magpie tool to do so

_____	_____	_____	_____
Trainee Signature	Date	Trainer Signature	Date

APPENDIX B HOW TO EXPORT FULCRUM DATA

This appendix describes how to export data from Fulcrum. You may need to export data for QC review or other reporting purposes.

a. **Exporting Data From the Record Editor**

i. From the application dashboard (***url***: https://web.fulcrumapp.com/), navigate and click on the "records" button of an application you wish to export data for. This brings you to what is called the "Record Editor" i.e. the location for entering records through a web browser.

ii. You can use custom filters to narrow down the subset of records you would like to export by following instructions on Fulcrum's website (link).

iii. You may also customize which columns are exported using the "Column Setup" button
NOTE: Only columns that are selected and viewable on this screen will be exported

iv. After custom filters are in place, if desired, click on the yellow "Download Data" button in the top right-hand corner of the Record Editor.

v. A new window will pop-up prompting you to select the desired file format
(1) Both .CSV and .XLSX file formats can be opened in Microsoft Excel

vi. Click "Start" after selecting a file format, the window will then say "Processing"
(1) This process will take more time with increasing numbers of records and child records

vii. A yellow "Download" button will appear in this window once processing completes, click this.

viii. The data will automatically save to the "Downloads" folder of your local computer.
(1) For applications with child and/or grandchild records, *the exported data will be divided into separate files* ("child" and "grandchild" are not keywords in the file names):
(a) One **parent** file
(b) One or more **child** files (if applicable)
(c) One or more **grandchild** files (if applicable)
(2) Rows across these files are related by several columns
(a) **Parent** files will have a single unique identifier column named "_record_id"
(b) **Child** and **grandchild** files will have three record identifier files "_child_record_id", "_record_id", and "_parent_id"
(i) In **Child** files, the "_record_id" and "_parent_id" columns will be equal to each other. Both of these fields correspond to the "_record_id" value in a parent file.
(ii) In **Grandchild** files, the "_record_id" value equals the "_record_id" in the parent file. Confusingly, the "_parent_id" equals the "_child_record_id" in the child file (because the child records are the direct "parent" of the grandchild record).
(3) Instructions for merging dozens of data columns across hundreds or thousands of child/grandchild records in Excel or other software are complex, error prone, and

beyond the scope of this document. If you require merged records, data exported from the Magpie applications are delivered as a single, fully merged file.

(4) Tutorials on how to use and manipulate data in Excel are provided in the Field Science Training Center (link).