NEON ALGORITHM THEORETICAL BASIS DOCUMENT (ATBD): OS GENERIC TRANSITIONS

PREPARED BY	ORGANIZATION	DATE
Sarah Elmendorf	DPS	01/05/2018
Hale Brownlee	CI	01/05/2018

APPROVALS	ORGANIZATION	APPROVAL DATE
Kate Thibault	SCI	02/14/2018
Mike Stewart	PM	02/19/2018

RELEASED BY	ORGANIZATION	RELEASE DATE
Judy Salazar	CM	02/21/2018

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	02/21/2018	ECO-05238	Initial Release

TABLE OF CONTENTS

1 DESCRIPTION **1**
1.1 Purpose . . . 1
1.2 Scope . . . 1

2 RELATED DOCUMENTS AND ACRONYMS **2**
2.1 Applicable Documents . . . 2
2.2 Reference Documents . . . 2
2.3 Acronyms . . . 2
2.4 Variables Reported . . . 3

3 SCIENTIFIC CONTEXT **4**
3.1 Theory of Algorithm . . . 4
3.2 Parser data constraints and validation . . . 4
3.2.1 Before checking for data ingest acceptance/failure . . . 4
3.2.2 Conditions for accepting/failing data ingest . . . 4
3.2.3 Conventions on dates and times . . . 5

4 ALGORITHM IMPLEMENTATION **6**
4.1 Summary of Algorithm . . . 6

LIST OF TABLES AND FIGURES

Figure 1 Overview of the transition system. The sample processing subroutine is outline in Figure 2; the Taxon Processing in Figures 3 and 4 . . . 7
Figure 2 Subroutine for processing samples. . . . 8
Figure 3 Subroutine for processing taxonomic data. Data may be entered either as a code (taxonID) or scientificName, depending on the product. When INPUT_TAXONOMY is specified in the publication workbook, original data are preserved except in cases involving rare, threatened and endangered species; when NEON_TAXONOMY is specified, taxonomic names are desynonymized and the higher taxonomy from NEON's taxon tables are provided. Master lists of NEON taxonomic names, codes, nativity and protected status, which have been assembled from a variety of published sources, can be found in NEON's Document library. . . . 9
Figure 4 Subroutine for taxonomic fuzzing. When a Federally or State-listed species is encountered, the taxonomic information is fuzzed to a higher taxonomic resolution. Where requested by site hosts, entire records about select taxa are redacted. . . . 10

1 DESCRIPTION

1.1 Purpose

This document details the algorithms used for creating a subset of NEON Level 1 data products that are the quality controlled products generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field, for example, the dry weights of litter functional groups from a single collection event are considered the lowest level (Level 0). Raw data are checked via front end validation, and then have limited ancillary data (as described here) appended to become Level 1 data products.

The text herein provides a detailed discussion of measurement theory and implementation, appropriate theoretical background, 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 theoretical background and entire algorithmic process for creating a subset of quality controlled and calibrated OS data products associated metadata from input data. It does not provide computational implementation details, except for cases where these stem directly from algorithmic choices explained here. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional details available in the NEON Ingest & Publications workbooks.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Applicable Documents

Applicable documents contain information that shall be applied in the current document. Examples are higher level requirements documents, standards, rules and regulations.

AD[01]	Nicl Language.pdf	NEON's Ingest Conversion Language (NICL) specifications
AD[02]	NEON.DOC.005003	NEON Scientific Data Products Catalog
AD[03]	NEON.DOC.002652	NEON Level 1, Level 2 and Level 3 Data Products Catalog

2.2 Reference Documents

Reference documents contain information complementing, explaining, detailing, or otherwise supporting the information included in the current document.

RD[01]	NEON.DOC.000008	NEON Acronym List
RD[02]	NEON.DOC.000243	NEON Glossary of Terms

2.3 Acronyms

Acronym	Definition
LOV	List of values
NICL	NEON's ingest conversion language
UTC	Coordinated Universal Time

2.4 Variables Reported

This ATBD describes the steps needed to generate OS L1 data products from the L0 products.

Some variables described in the relevant publication workbooks may be for NEON internal use only and will not appear in downloaded data. These are indicated with **downloadPkg** = “none” in the publication workbooks.

3 SCIENTIFIC CONTEXT

3.1 Theory of Algorithm

This document describes the algorithms for assessing the integrity of the L0 data stream generated by the field sampling and lab analysis of OS data.

The approaches described herein are simple yet necessary components of quality control and quality assurance, and include such processing steps as: verifying that all required data are recorded for each observation event, comparing data values to pre-defined (through provided validation rules or specified lookup tables) ranges of possible values, and tracking the individual-level data for consistency and accuracy through time.

3.2 Parser data constraints and validation

Many quality control measures are implemented at the point of data entry at the time data are uploaded, either when the contents of the data entry forms (specific to each product), are parsed into the database, or when a user uploads a .csv or other file type containing data to be ingested. Rules for such quality control measures are defined in the ingest workbooks, such as the **dataType** for each field and which fields are required. Constraining data formats reduces the number of processing steps necessary to prepare the raw data for publication. This section describes the data constraint and validation requirements that are built into the ingest process.

3.2.1 Before checking for data ingest acceptance/failure

1. Strip any lines that contain only empty values from file
2. Strip all starting and ending double quotes and convert inner quotes to single quotes
3. Replace all non-leading/trailing line endings, tabs, ctr-R in all string fields with a single space
4. Strip all leading and trailing white space, including line endings, tabs, and ctr-R

3.2.2 Conditions for accepting/failing data ingest

If there are errors on upload, return the following information to the user: row number; field name; brief description of what failed. If multiple locations in the file fail, print all messages.

1. Reject upload if any value exceeds 4000 bytes
2. Reject if **dataType** does not match that specified in the publication workbook. For numeric fields:'Integer' passes if numeric values are integer or if all digits after the point are zero (e.g., 3 passes and so does 3.000)
3. Validation rules are specified in the ingest workbook **entryValidationRulesParser**
4. activityEndDate must always be later than or equal to activityStartDate
5. All dates must fall between 2010 and the systemDateTime + 24 hours
6. For validation of type [LOV] (lists of values), match entered values to the lovElementName, in a case-insensitive manner. Post the corresponding lovElementCode to the database, in the case specified in the LOV
7. The parser will not validate data from LOVs that are entered using forms; only for files directly uploaded to the parser

8. All entry validation rules (except REQUIRE) imply if not null then do the rest of the validation
9. All date fields can be entered as dates or dateTimes, the parser will interpret whether time is included based on the formatting
10. Specifics on machine-read validation rules can be found in AD[01], NEON's ingest conversion language

3.2.3 Conventions on dates and times

1. Incoming dates are interpreted at UTC, unless conversion from local dates is specified in the ingest workbook
2. Dates without times are interpreted as noon local time

4 ALGORITHM IMPLEMENTATION

Data are transitioned in the NEON database according to the wait and search interval parameters of the OS transition system. Lag dates from data collection to automated processing vary by product and table within product, in order to account for anticipated durations of data entry and qa by field staff, laboratory procedures and shipping times. Lags range from several weeks to 18 months.

4.1 Summary of Algorithm

Algorithm implementation is outlined in the following figures.

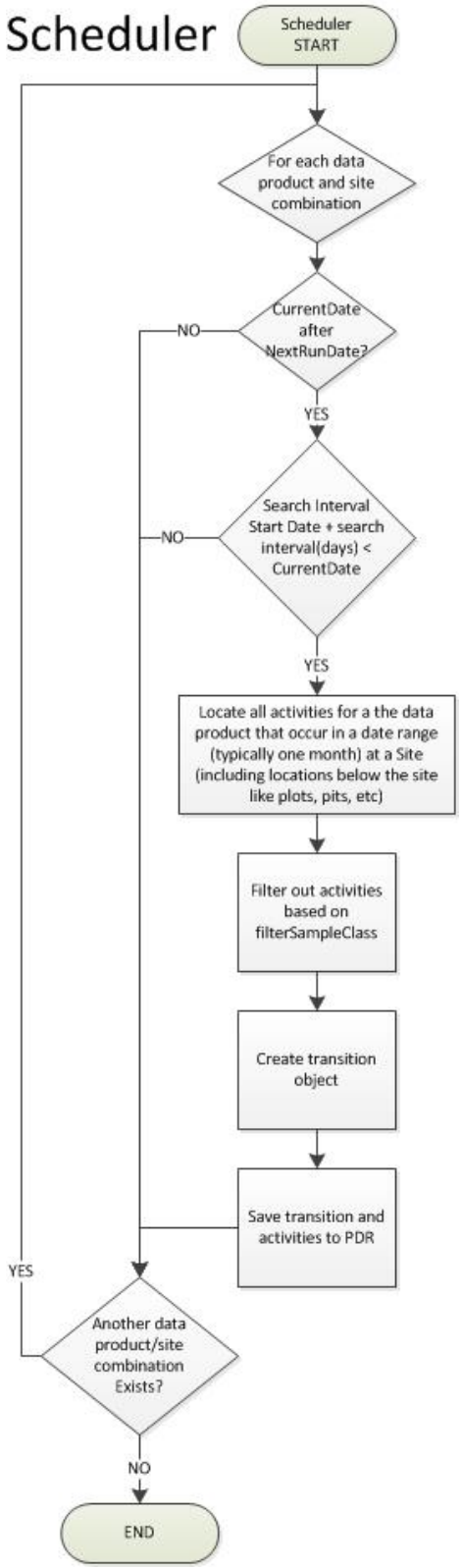

Figure 1: Overview of the transition system. The sample processing subroutine is outline in Figure 2; the Taxon Processing in Figures 3 and 4

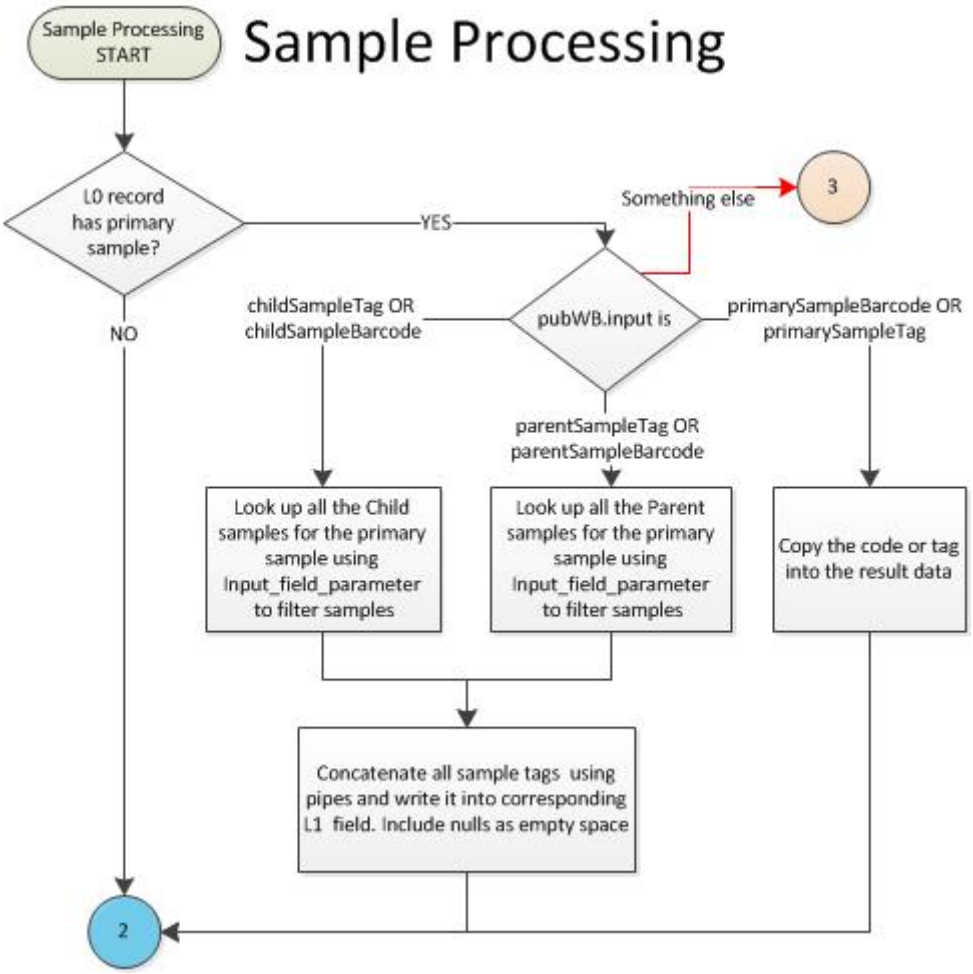

Figure 2: Subroutine for processing samples.

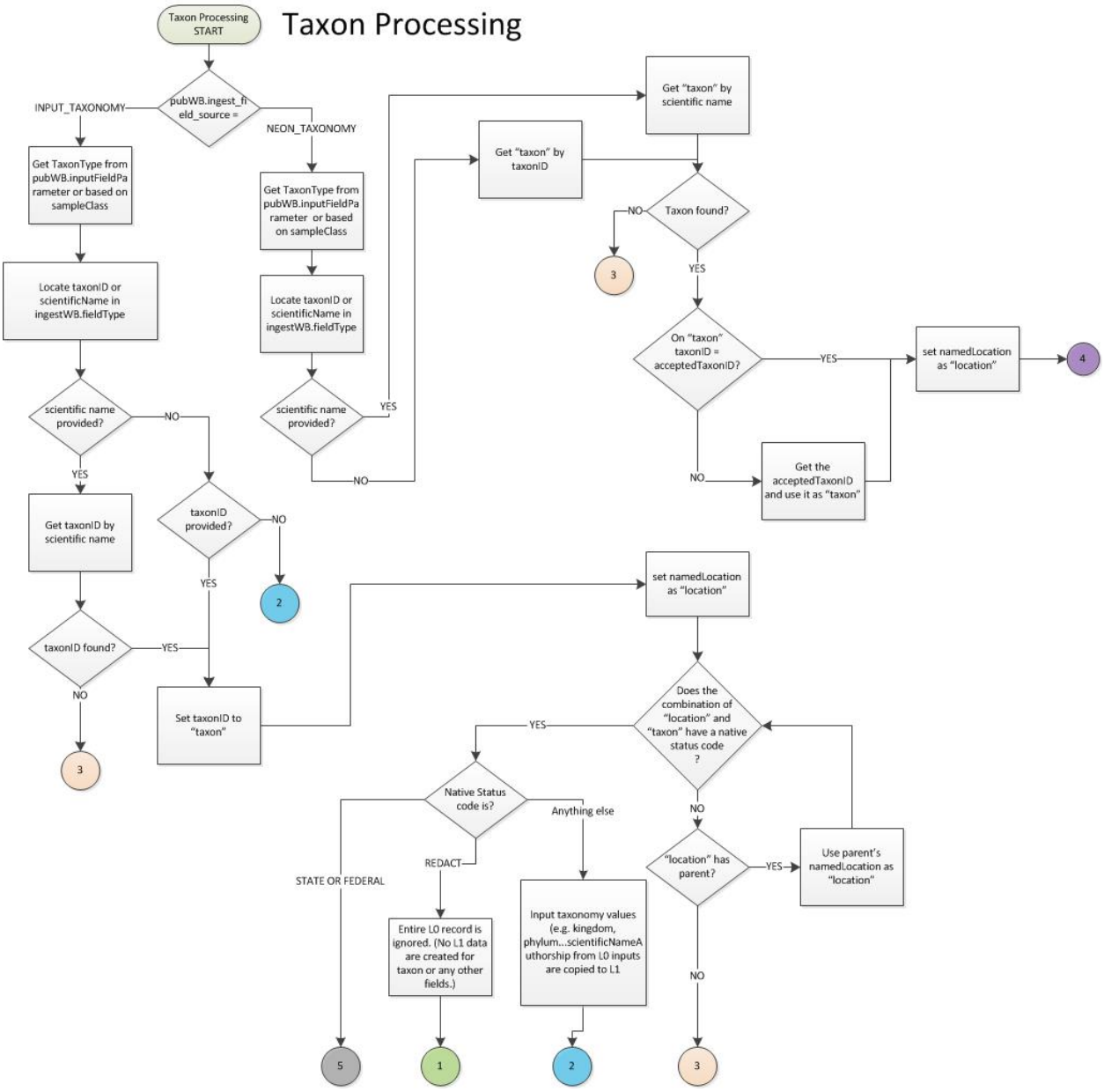

Figure 3: Subroutine for processing taxonomic data. Data may be entered either as a code (taxonID) or scientificName, depending on the product. When INPUT_TAXONOMY is specified in the publication workbook, original data are preserved except in cases involving rare, threatened and endangered species; when NEON_TAXONOMY is specified, taxonomic names are desynonymized and the higher taxonomy from NEON's taxon tables are provided. Master lists of NEON taxonomic names, codes, nativity and protected status, which have been assembled from a variety of published sources, can be found in NEON's Document library.

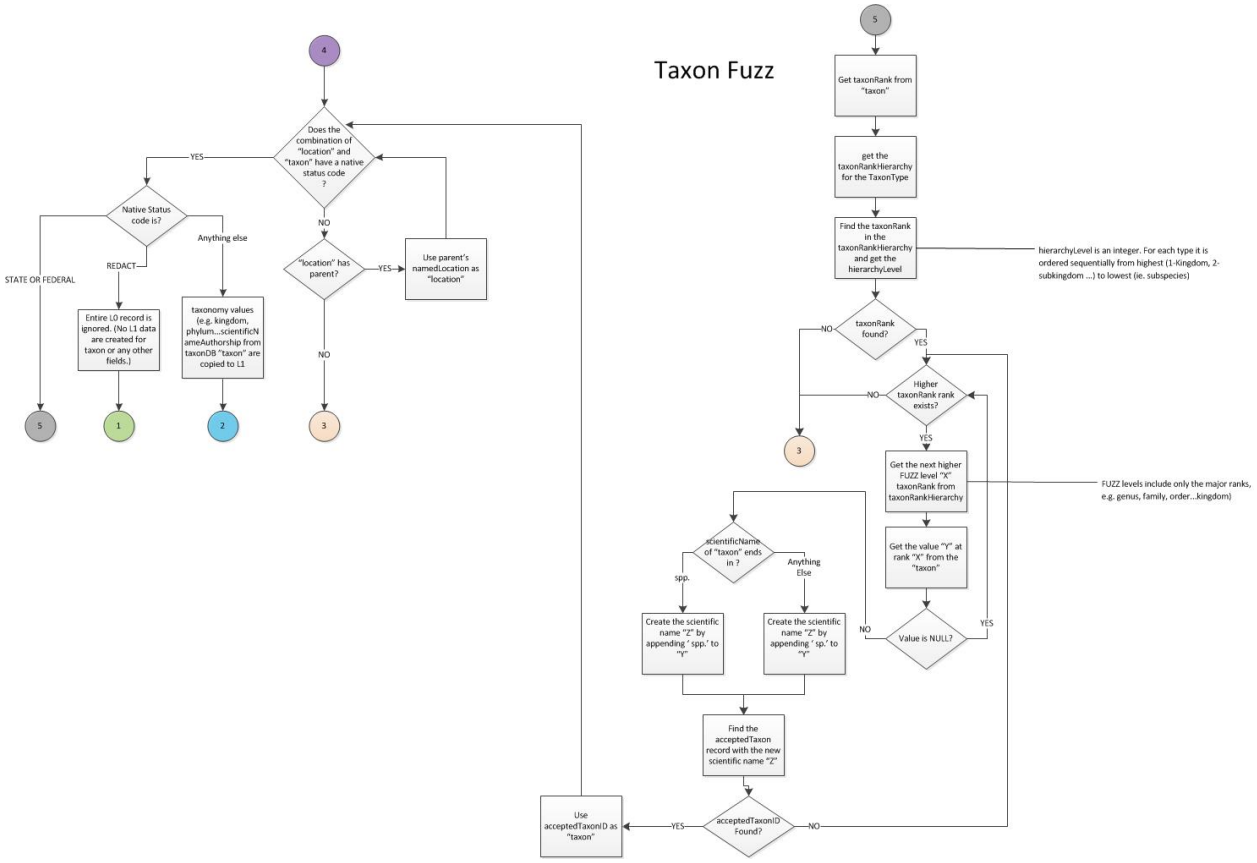

Figure 4: Subroutine for taxonomic fuzzing. When a Federally or State-listed species is encountered, the taxonomic information is fuzzed to a higher taxonomic resolution. Where requested by site hosts, entire records about select taxa are redacted.