**StreamNet Data-Contributing Partners' QA/QC Procedures**

Before data is shared with StreamNet, it first undergoes QA/QC procedures at the source. What follows is a summary of those processes for our current StreamNet partners. April, 2019

Confederated Tribes of the Colville Reservation

The Colville Tribes submit data to StreamNet that flow from field to StreamNet in four steps: field collection by Colville Tribes biologists; data aggregation and calculations; importing data to the Colville Tribes database; and importing data to the StreamNet database. Quality assurance (QA) is ensured at the field collection stage by using well established protocols, standardized training of field technicians, and, where possible, the use of electronic data collection devices (e.g. P4 for PIT tagging) that use allowable value constraints, pick lists, and required fields. All field data are collected by at least 2 crew members to ensure accuracy, e.g. when PIT tagging juveniles the crew member entering data confirms fish measurements and species identification called out by the tagger.

To ensure quality control (QC), field data are reviewed daily and uploaded where appropriate (P4), allowing biologists to isolate any data aberrations to a single field day. All data entered manually are reviewed by a second crew member. Geographic data are post-processed and checked for out-of-bounds features. When possible, data are checked for consistency by comparing values to previous data collection events.

Post-aggregation/calculation spreadsheets and shape files are sent to Sitka Technology Group to store in a single database where a second level of QA/QC processes are applied. Tables contain allowable value constraints, spreadsheets are checked for outliers via assessing value ranges or producing exploratory plots, and calculations embedded in data spreadsheets are checked for accuracy. Data are uploaded to StreamNet automatically using a Python script.

Idaho Department of Fish and Game

The Idaho Department of Fish and Game (IDFG) submits data to StreamNet from several different programs within IDFG (research, regional management, hatchery production), and from tribal and federal collaborators. Data are extracted, transformed, and loaded (ETL) from their sources, validated against the StreamNet (SN) and Coordinated Assessments (CA) Data Exchange Structure (DES), and then uploaded via a number of application programming interfaces (API).

The quality assurance (QA) and quality control (QC) processes vary by data source and level of implementation. Compliance with IDFG QAQC processes is mostly voluntary with our collaborators. A common strategy is to start with the current year, and work backward as time and resources allow. Feedback from users about possible errors or structural inadequacies is evaluated and shared with collaborators before implementation.

In general, there are two levels of QAQC processes which occur at four different stages between the collection of data in the field and its submittal to SN. QAQC procedures occur at both the user level and the programmatic level during the data collection, entry, compilation, and submittal stages.

User level QA processes occur during the training, collection, and data entry stages. They include standardized training and user manuals for common, standardized data entry applications (e.g. hydrography, taxonomy). User level QC includes comparing field records with the data in local databases, and then comparisons between local databases and central databases for completeness and correctness.

Programmatic level QAQC processes occur during all stages. Data entry typically occurs via the use of IDFG data entry applications in which column and row constraints are applied, including value ranges, pick lists, required fields, predefined formats, validation rules for geographic location and temporal extent, logical consistency between fields and tables (e.g. primary key fields, join definitions), orphan queries between parent and child tables, and duplicate queries within tables.

Data prepared for SN and CA submittals are queried from internal databases or compiled from documents or digital files. The data are manipulated via queries into the SN or CA data exchange structures (DES). The data are then georeferenced with the SN hydrography, checked for logical consistency (orphans, duplicates), and then appended into an IDFG SQL database. Once the data have been appended into the local SN database, they are validated against SN and CA DES rules, and uploaded to PSMFC databases via the IDFG StreamNet API or the Coordinated Assessments API.

The data undergo further QAQC processing by the SN database manager. Any necessary corrections are made at the state level and then resubmitted to the SN database manager before posting on the SN website. The data and references are then compared between the SN website and the local SN database for completeness and correctness (QC). This is done via the StreamNet Exchange (StreamNetX) API. All corrections are made at the source and resubmitted, or coordinated between the IDFG and the PSMFC version. This is a very efficient, linear data flow when the ETL and QAQC processes are dynamic links (e.g. queries and views) between the data sources, the local StreamNet SQL database, and the regional SQL database.

Two exceptions to this data flow model are the Generalized Fish Distribution (GFD) and the Columbia River Basin Fish Facilities (CRBFF). They are primarily utilized as geographic information system (GIS) data, and are compiled, managed, and shared as GIS data. The same QAQC procedures apply to their tabular attribute data, but those are relatively limited and simple flat files. The geometric data (points, lines, polygons) are limited to a shared hydrography or waterbody network with a standard datum and projection for quality assurance. Furthermore, IDFG data are collected in a standardized datum and projection. This makes the ETL of these data from IDFG sources into the SN DES and submittal to SN GIS efficient and introduces less error during reprojection. The GIS layers from IDFG and SN are easily compared for completeness and correctness (quality control).

Montana Fish, Wildlife & Parks

Montana Fish Wildlife & Parks (MFWP) has implemented QA/QC procedures for data submitted to StreamNet in several ways. There are two general levels of QA procedures that occur at three different stages between the collection of data in the field and its submittal to StreamNet. Quality assurance procedures occur at both the user level and the programmatic level during the data collection, entry, compilation, and submittal stages.

User level QA occurs during the collection of data and upon its subsequent entry into the MFWP centralized database, the Fisheries Information System (FIS). Training and user guides have been created and given to field staff to assist in consistent data entry. All Fisheries staff use the same field survey forms when conducting netting, trawl, trapping, and electrofishing surveys. Data entered by field staff into the agency’s centralized internal database is by default marked as provisional until it has been checked for accuracy. Once data has under gone biologist or supervisor QC it is marked as reviewed. Only reviewed data can be analyzed in the system or downloaded for other uses.

After data are entered into the MFWP internal database, several layers of QC exist. Data entry fields have constraints that ensure consistency including value ranges, pick lists, required fields, predefined formats, validation rules for geographic location, logical consistency between fields, orphan queries between parent and child tables, and prevention of duplicate records.

Data submitted to StreamNet from FIS undergoes additional QA/QC. The data are transformed by a series of queries into a consistent format, then transferred to a secondary internal database. Agency StreamNet staff use a multi-part python script to further QA/QC this dataset and ensure that the consistent procedures are used. The script performs a number of validation queries, compiles data into the necessary exchange table format, then transforms data to meet requirements of the StreamNet DEF. The data is also georeferenced with the regional multiscale hydrography, checked for logical consistency, orphans, and duplicates, and then appended into a local StreamNet database. Upon completion, data are submitted to the StreamNet API where StreamNet preforms real-time record level validation before accepting data.

Oregon Department of Fish and Wildlife

The Oregon Department of Fish and Wildlife (ODFW) utilizes staff within the Natural Resource Information Management Program (NRIMP) to request and gather reports, databases and other data sets from around the state. NRIMP is responsible for entering fisheries data into MS Access databases that align with StreamNet data exchange standards (DES), as well as validation rules developed for the StreamNet API. Data are obtained in databases and spreadsheets upon confirmation by district biologists or project leaders that the data are ready for sharing. In addition to abiding by StreamNet’s guidelines for data entry and QA/QC (https://www.streamnet.org/data/coordinated-assessments/data-qa-qc/), ODFW follows other specific procedures described below. Oregon StreamNet staff continually evaluate current systems and processes to ensure they are the most efficient and effective possible given ODFW’s data management capabilities.

ODFW StreamNet staff apply QC processes during data entry and prior to submission to StreamNet. Updated records (between submissions) are QC’d by comparing them to historic and current references as appropriate. Data entry errors are corrected at the time they are identified and records are marked for update, corresponding to the next submission. Data providers are notified of errors found in data they submit, with a request to resubmit corrected information for archiving. Depending on time and resources, no less than 10% of all records maintained in the database are randomly selected annually for a manual QA review by someone other than the person who did the original data entry. Survey locations on streams are compared against whole stream route events derived from the USGS National Hydrography Dataset (NHD) for verification. When new or updated coordinates are available, GIS linear referencing is conducted to accurately identify and capture current sampling locations.

Following the current DES developed by StreamNet and agency partners, validation rules are applied and checks run on all required fields to maintain consistent and accurate information within the database. ODFW staff utilize electronic form-based data entry screens, with instant validation of acceptable values via look-up tables. The ODFW database also contains several QA/QC queries that are run periodically throughout the year on data entered since the last submission to StreamNet. These queries are used to identify and resolve potential errors arising during data entry or QA/QC efforts, including calculation errors, appropriate use of DES codes, spatial accuracy, etc.

The submission process to StreamNet calls DES tables from the Access database to be imported into a SQL Server database. Several checks take place during this process including duplicate primary keys, missing references for foreign keys, the existence of values for required fields, and proper format of all fields to ensure completeness of the data, and that new records satisfy all validation rule requirements. When the data are deemed acceptable, they are submitted to the central StreamNet database via the StreamNet API, where they are validated a final time before posting to the StreamNet online query system. During the submission process, ODFW’s system automatically reads each individual submitted record back to validate it transmitted correctly, and compares table record totals with StreamNet’s system to ensure the number of records match.

Washington Department of Fish and Wildlife

1. WDFW StreamNet Data Stewards have built and are building regional and corporate databases that house and are able to report on collected data. These relational databases enforce referential integrity rules using drop-down pick lists, validation rules, and standardized input formats to control what is entered.
2. WDFW StreamNet data mangers receive data from either corporate data sources like FishBooks (Hatchery Returns), Spawner Ground Survey (SGS), regional data sources, Age and Scales (Biological data), Traps/Weirs/Surveys (Mark Recapture), and the Sport and Commercial (Mainstem Columbia River sport and commercial data) databases, or from the regional biologist responsible for the final product. Data from the regional biologists can come in many different forms. These include, but are not limited to; memos, electronic files (Excel, MS Word, Access, GIS, ASCII), finalized agency reports, or personal communication by phone or email. Data received via personal communication are documented in a memo and sent to the original contact for approval. Data received via an electronic database are referenced to a published report if a report is ultimately published to validate the original data.
3. When data meet the minimum requirements for the StreamNet DES, the Data Stewards either enter the final data or transfers the data via electronic routines into the appropriate internal WDFW StreamNet database. The internal database includes many fields that are useful for internal tracking and quality control that are too agency-specific or cumbersome to include in StreamNet’s DES. The internal WDFW StreamNet databases have a data entry interface or stored procedures for transferring electronic data. The databases use drop-down pick lists, validate the entries within prescribed limits and enforce referential integrity (rules established between tables).

If data DO NOT meet the minimum requirements of the DES, the compiler contacts the source of the data to get complete information. Yet, sometimes the biologist is too busy to finalize the estimates. If the WDFW StreamNet Data Steward is equipped, the Data Steward applies expansion factors or runs the raw data through known models to finalize an estimate and sends it out for final approval from the biologist.

After any data have been entered or pulled into the appropriate WDFW StreamNet database, the compilers verify the database version against the original data. This process is also repeated for previous years data as updates to the data by the biologists are sometimes made and sometimes preliminary numbers are submitted and need to be proofed once a final number is approved.

Procedures enacted in each data partner (Biologist, Region Database, or Corporate Database) must ensure smooth data flow from collection to StreamNet data exchange and must be carefully designed. Those procedures must include competent data validity checks to fill data gaps and correct erroneous values without introducing additional error through misinterpretation or accidental data manipulation errors.

1. When data are to be submitted to StreamNet, the compiler pulls a snapshot of the data from the database. The compiler checks the snapshot for completeness and accuracy using various routines and documents the purpose of each table and any noteworthy issues. The WDFW Data Submitter further checks the data, making any final conversions that are warranted per StreamNet’s DES. References are sent to the SN Librarian.
2. Any identified data errors are sent back to the originator to correct the original data set.