**Summary of quality control procedures**

**used to merge agency data with StreamNet’s regional data**

**Bill Kinney**

**StreamNet Database Manager**

**March 24, 2012 Draft**

Fisheries data compiled in state, federal and tribal agencies and sent to the regional StreamNet project at PSMFC are predominantly transferred in a compressed MS Access database of tables defined by the current StreamNet data exchange format ([DEF](http://www.streamnet.org/def.html)), and are accompanied by a data exchange form (Word document) that includes a summary of the information sent, instructions related to any caveats or special considerations necessary for processing (such as orphaned record keys that need to be deleted) and any detailed notes that the agency wants associated with the data submission regarding the preparation methods used or the expected delivery of future data updates. If new references are cited in the data submission, a spreadsheet is included that describes the new documents, and that spreadsheet and a physical or scanned copy of the document is sent to the StreamNet Library for processing. Upon receipt of the data submissions, a series of procedures is initiated to ensure that the data fully comply with the DEF and will load correctly into the StreamNet database. This document outlines those procedures and how they are applied. These actions are in addition to QA procedures applied at the agency level that occur prior to submitting the data to the StreamNet office at PSMFC (Appendices A through E).

Data submissions are logged and stored on a data server in a hierarchical system by agency, year, and a unique date-specific folder for each submission. The regional data manager extracts the documents and immediately verifies that the data submission is intact and uncorrupted. After reading the data exchange form document and any email that accompanied the submission, the database is opened and the appropriate macro or macros needed to process the data categories submitted are imported into the database from a code library database.

The code library database contains over 1000 objects, consisting of about 800 queries, 100 linked SQL Server tables, 80 mostly empty local tables matching the StreamNet DEF, 19 macros customized to perform various actions, and a few forms. Macros contain unlimited numbers of actions, and most are made up of up to a hundred transfer-database commands that link master tables and import queries into the submitted database. When the macro that is customized to process the given data set is run inside of a newly submitted database, it will link all appropriate master SQL Server tables related to the categories of data submitted (both the tables that need updating and any related look-up tables), and will import the appropriate series of sequentially-numbered queries to process the data, a table designed to track progress and any issues encountered while processing the submission, and any additional associated macros into the working database that may be used to execute numerous queries at once.

The validation queries systematically check each table for valid data in all required fields, by either performing an outer join (which will list any values not included in related look-up tables) or by using SQL (Structured Query Language) syntax in a query that lists any values in a field that are not within a list or range defined in the DEF. For example, data records must include LocationIDs found in the regional mixed-scale hydrography, and RefIDs (codes relating to reference citations) must exist in the StreamNet Library catalog. Any tables missing valid information in required fields or containing suspicious changes to existing primary keys (without an expressed request for deletion) are referred back to the agency compiler for correction, usually accompanied by a table pasted into a Word document containing the data in question, often with probable solutions to consider. Some queries test for logical consistency errors between tables, e.g. when data are attributed to stream reach measures beyond the length of a stream, and mathematical errors within tables, e.g. when various columns should add up to a totals column. Additional queries test for referential integrity constraint issues with related tables prior to insertion of new data; i.e. record primary key values must be unique within each table, and any values which are required to be defined in look-up tables must already exist in those tables. Detailed notes are recorded during the validation process, and are later filed in the data folder containing both original and final versions of working databases, along with all related email correspondence and documentation.

Once submitted data are validated, new records are appended to StreamNet’s MS SQL Server (Microsoft relational database program) tables. Then, a sequential series of queries is run to compare the field values of all submitted Access tables with those of the SQL Server tables and differences that are visually approved by the regional data manager are made to the primary regional database by either executing update queries or manually pasting the new column data into appropriate rows. A SQL Server trigger is fired after any new record is added or existing record is updated to track the person making the change and the date each record is modified.

After data submissions are synchronized with the regional database, another batch of SQL scripts and stored procedures is run on the entire StreamNet database to detect possible issues such as records orphaned by primary key changes, identify records that may overlap other agency data, and to update tracking fields such as those that enumerate the occurrences of LocationIDs and RefIDs used in trend and count data records, as well as to assure that any new look-up table values are added to specialized tables that facilitate retrieval of data through the StreamNet web-based query system (WQS). Any queries that were modified or added to the processing steps are replaced or added in the code library database, and the macros are edited to account for any new steps (queries) that need to be part of processing the next similar data submission (hence, the reason why the library database must have links to back-end SQL Server tables and dummy local tables; Access would not import queries successfully if the objects they use do not exist).

When data are completely validated in the master database, they are moved to the production database that serves the WQS and to a StreamNet\_X SQL Server database. StreamNet\_X is a copy of the main database that is accessible to compilers outside the PSMFC firewall via SQL Server management Studio or via Access ODBC connection that allows project compilers in the data source agencies to compare their data with the regional data and to pre-validate new data prior to submitting. An Access database containing most StreamNet tables and a few sample queries is also made available about twice each year through our website for anyone who would like to obtain the entire StreamNet database. The detailed procedures used to update the master database and port the database and updated binary files to the production and external servers are maintained by the data manager.

Periodically, additional steps are needed to assure quality control of StreamNet data. For example, when a new version of mixed-scale hydrography is adopted by StreamNet, location coding and reach measures in tabular data may need to be adjusted and verified, and geographic referencing look-up tables that allow filtering by the WQS need to be regenerated. Another example is when the DEF is modified or additional functionality in the WQS requires additional fields; under those circumstances all existing table data must be updated to be compliant. Of course, regular database management duties such as scheduling and testing routine backups, rebuilding indexes, auditing server logs, and performance tuning are also done by the data manager.

The regional staff hosts quarterly meetings with StreamNet’s data compilers and GIS technicians in the data source agencies to address data standards and consistency issues, explain new techniques, methods or software being used, share best-practices and resolve problems. More often, issues regarding questionable data that are encountered by StreamNet regional staff or brought to our attention by users through various feedback alternatives on the StreamNet website are forwarded to data compilers in the data source agencies for verification or correction and resubmission via email. Ideas regarding new or better methods to improve StreamNet data exchange formats are typically handled by a an online Forum that tracks and archives the issue, opinions, suggestions and conclusions that are presented for approval by our Steering Committee prior to adoption.

Appendix A

**Summary of quality assurance and quality control (QAQC) procedures**

**Employed by Montana Fish, Wildlife & Parks fisheries data submitted to StreamNet**

**February 29, 2012**

Montana Fish, Wildlife & Parks takes data QC very seriously. Fisheries biologists and technicians QC data during and after collection to make certain that data transmitted to MFWP databases are accurate. QC that takes place at the biologist/technician level primarily consists of location and fish species accuracy. Sample points collected by GPS units are plotted using GIS software to verify they are in the correct location. Species recorded on data sheets are verified to be certain that species recorded are within their distribution range areas. Fish population estimates are calculated either in a database or using spreadsheet software to minimize errors in computation. Biologists review population estimates derived for accuracy before committing them to a database and before they are sent to MFWP StreamNet staff.

Montana FWP StreamNet staff employs data QC processes both at the time of data entry and before submission to StreamNet. Internal agency database QC consists of allowed value constraints to ensure compatibility with StreamNet lookup values. On tables where allowed value constraints are impractical, queries are run to find and update non allowed values. Distribution data are queried against survey and inventory data to make sure that all entries to survey and inventory data are sufficiently represented in species distribution. All data are checked to make sure that they fall within the extent of the associated stream route. These are just a few descriptions of the 20 QC queries that are run against the data in the internal MFWP database before the data are prepped for submission to StreamNet.

Prior to data submission to StreamNet, all data are verified against the mixed scale hydrography to ensure that the associated hydrography is represented and that points and routes are in the appropriate locations. Data are also verified against the StreamNet database using an ODBC connection. Queries developed by PSMFC StreamNet staff are run on the data as a final check before data are submitted.

Appendix B

**WDFW StreamNet Staff QA/QC Methodologies**

**March 9, 2012**

1. WDFW StreamNet data stewards 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 (Main stem 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.

WDFW StreamNet Data Stewards have built and are building databases that will house collected data until corporate databases are built to house this raw data. These relational databases enforce referential integrity rules using drop-down pick lists, validation rules, and standardized input formats to control what is entered.

1. When data meet the minimum requirements for the StreamNet DEF, 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 DEF. 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 DEF, 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 DEF.

Data transferred to the regional StreamNet data manager are accompanied by a data exchange form that lists what is being sent and any special instructions applicable to the data. The data are then run through programmed routines that check for errors and completeness before being accepted. Any data in question are noted and sent back to the source for verification.

References and a reference submittal document are sent to the SN Librarian.

1. If StreamNet’s Regional Data Manager has any issues with the submission, the Regional Data Manager and the data submitter discuss the issues and determine the best process for resolution. Any identified data errors are sent back to the originator to correct the original data set.

Appendix C

**Quality Assurance Process Applied to StreamNet Data at the Idaho Department of Fish and Game**

**March 24, 2012**

The Idaho Department of Fish and Game (IDFG) submits data to StreamNet from several different programs within IDFG. The quality assurance (QA) processes vary by data source. In general, there are two levels of QA processes which occur at four 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 processes occur during the collection and data entry stages. They include standardized training and user manuals for common, standardized data entry applications, comparing field records with local databases and comparisons between local databases and central databases.

Programmatic level processes occur during all stages. Data entry typically occurs via the use of IDFG data entry applications in which field constraints are applied, 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 duplicate records.

Data prepared for StreamNet submittals are queried from internal databases or compiled from documents. The data are manipulated via queries into the StreamNet data exchange format (DEF). The data are then georeferenced with the StreamNet hydrography, checked for logical consistency, orphans, and duplicates, and then appended into a local StreamNet database. Once the data have been appended into the local StreamNet database, they are packaged along with a data submittal summary and any relevant reference documents and sent to PSMFC

The data undergo further QA processing by the StreamNet database manager. Any necessary corrections are made at the state level and then resubmitted to the StreamNet database manager before posting on the StreamNet website. The data and references are then compared between the StreamNet website and the local StreamNet database.

Appendix D

QA Process Applied to StreamNet Data at US Fish and Wildlife Service

Appendix E

QA Process Applied to StreamNet Data at Oregon Department of Fish and Wildlife