**StreamNet Quality Assurance and Quality Control Plan\***

**Version – February 7, 2022**

\*This plan should be considered a “living document” as updates will occur as needed and through input received during annual technical workshops.

Table of Contents

[**Phase 1 - Quality Assurance** 4](#_Toc95658137)

[QA Procedures 4](#_Toc95658138)

[Fields Requiring Special Attention 5](#_Toc95658139)

[**Phase 2 - Quality Control** 5](#_Toc95658140)

[QC Procedures 6](#_Toc95658141)

[Validation Rules 6](#_Toc95658142)

[Visual Checks 6](#_Toc95658143)

[Timestamps 8](#_Toc95658144)

[Fields Requiring Special Attention 9](#_Toc95658145)

[**Phase 3 - Quality Data Visualization** 9](#_Toc95658146)

[Effectively Communicating to Data Consumers 10](#_Toc95658147)

[Fields Requiring Special Attention 10](#_Toc95658148)

[**Issues Beyond QA/QC Data Flow Tasks** 11](#_Toc95658149)

[**Conclusion** 13](#_Toc95658150)

[**Appendix A: 2019 StreamNet Funded Data-Contributing Partners' QA/QC Procedures** 14](#_Toc95658151)

[Confederated Tribes of the Colville Reservation 14](#_Toc95658152)

[Idaho Department of Fish and Game 14](#_Toc95658153)

[Montana Fish, Wildlife & Parks 15](#_Toc95658154)

[Oregon Department of Fish and Wildlife 16](#_Toc95658155)

[Washington Department of Fish and Wildlife 17](#_Toc95658156)

[**Appendix B: Summary of 2021 interviews with Data Providers and Consumers** 18](#_Toc95658157)

[Data Consumers 18](#_Toc95658158)

[Data Stewards 21](#_Toc95658159)

[**Appendix C: Manual Review of CAP Fish HLIs Data to Discover Issues, Assess, and Propose Solutions** 27](#_Toc95658160)

[Approach 27](#_Toc95658161)

[Records Reviewed 27](#_Toc95658162)

[Summary of Issues Discovered 28](#_Toc95658163)

[Problematic Fields 28](#_Toc95658164)

[Recommended Solutions and Existing Group 30](#_Toc95658165)

[**Appendix D: Boilerplate Template for Data Stewards to Communicate to Biologists the Expected Content for Problematic Fields** 38](#_Toc95658166)

**Background**

The Pacific States Marine Fisheries Commission’s (PSMFC) StreamNet Program (StreamNet) provides access to standardized data to inform regional assessments, reports, and decision-making processes. This necessitates that quality assurance (QA) and quality control (QC) standards are met, while the best available information is clearly communicated. The StreamNet Quality Assurance and Quality Control Plan was developed to ensure that the quality of the data satisfies the needs of the National Oceanic and Atmospheric Administration Fisheries (NOAA), Bonneville Power Administration (BPA), Northwest Power and Conservation Council (NPCC), U.S. Fish and Wildlife Service (USFWS), and other data consumers.

Data submitted to StreamNet's Coordinated Assessment Partnership (CAP) Fish High-Level Indicators (HLI) and Fish Monitoring Data originate from fish and wildlife agencies and tribes. These entities implement pre- and post-submittal QA/QC processes that evaluate data and resolve existing issues. Documenting these processes contribute to data quality and integrity, which provides data consumers with the confidence to use the information, as well as facilitating the resolution of future questions regarding the data. To communicate the QA/QC implemented by data providers and PSMFC-StreamNet staff, this plan partitions the effort into the following phases:

***Phase 1 - Quality Assurance***: Ensure exchanged data sets are the best available information and are submitted per data exchange standards. This includes work performed by biologists for data collection, entry, analysis, and metadata documentation, as well as efforts by data stewards to error-check and convert data and metadata into the appropriate data exchange standards (DESs).

***Phase 2 -*** ***Quality Control***: Verify exchanged data sets are complete, correct, and meet DESs. This phase includes automated validation rules that verify data upon submittal to the central data system, as well as work performed post-submittal by the data providers, data stewards, independent reviewers, PSMFC-StreamNet staff, and data consumers. Also included in this phase is the documentation of completed quality control checks and the tracking of changes that occur post-submittal.

***Phase 3 -*** ***Quality Data Visualization:*** Ensure that data displayed and downloaded are communicated correctly and effectively to support use by data consumers. This focuses on how StreamNet products and tools communicate and distribute information to data consumers (e.g., online queries, downloadable/exported files and spatial data products such as web map services, mapping applications and GIS data download packages).

The *Issues Beyond QA/QC Data Flow Tasks* section addresses aspects that arose while developing this plan, that contribute to the data quality and data integrity but that fall beyond the QA/QC data flow tasks. For instance, one issue that arose related to QA/QC is the need to have a strong collaboration between biologists and data stewards, built on a common understanding of the correct content to be submitted for each field. In this section these issues are summarized, as these cannot be addressed through the QA/QC Plan. Existing groups have been identified that could assess and rectify the underlying causes.

The foundation for this plan, which was developed to ensure consistent QA/QC practices, is based on information provided by data providers and data consumers of the CAP Fish HLIs (CAX), as well as reviews of the CAP Fish HLIs (Appendices A and B). A manual review of the CAP Fish HLIs data was conducted to assess issues identified by data consumers and to determine whether additional issues exist (Appendix C).

## **Phase 1 - Quality Assurance**

Quality Assurance (QA), for the purpose of this plan, refers to the process of ensuring that the data exchanged result in the correct data informing the StreamNet database. Quality assurance focuses on prevention of errors and omissions. Typical quality assurance tools include check lists, data cross-checks, and documented standards. QA activities typically occur up-front in a project (modified from [2011 DAMA Dictionary](https://www.dama.org/cpages/body-of-knowledge)). In performing quality assurance data stewards submitting information to the CAX complete data validity checks to fill data gaps and correct erroneous values, while avoiding the introduction of additional error through misinterpretation or inadvertent data manipulation. Quality assurance implemented within the data owner's/provider's organization should be applied to data collection, entry, compilation, and submittal to a local data system (if applicable), prior to submitting to StreamNet.

### QA Procedures

To ensure a minimum level of quality assurance is performed by all data stewards, prior to submitting data to StreamNet, the following steps should be implemented during data collection, entry, and compilation:

* For manual data entry, double-check data entry (e.g., enter data twice, check a random sample of data, have another staff check the data entry, have multiple staff enter the data to facilitate comparisons and the identification of errors). For distributed data systems with multiple biologists and technicians entering data into the data system a variety of data entry verification approaches may be in use.
* Compare new data records with historic and current references to identify data aberrations.
* If data entry errors are detected, for either previously submitted records or new records being prepared for submittal, identify the source (e.g., typo, statistical) and notify relevant staff.
* As feasible, use field constraints to govern values entered in a field and limit inconsistencies and erroneous entries. Data entry should leverage automated validation rules to reduce human error by applying rules for format, length, range, required fields, acceptable values, drop-down pick lists, orphan records, and duplicate records.
* If appropriate, data stewards should apply expansion factors or process raw data, through models, to obtain an estimate for the CAP Fish HLIs that would subsequently require confirmation by a biologist.
* After data are reviewed by the appropriate staff (e.g., a project leader, field biologist, district biologist, or supervisor) and determined to meet quality assurance standards, the information should be prepared for submittal to StreamNet.
* Compare field records to data in local databases and between local and central databases to verify completeness (i.e., no missing values) and correctness (i.e., correct field, format, and spelling).
* Depending on resources, a percentage of all records should be visually reviewed, on an annual basis, by someone other than the person that entered the data.
* Consider using training and user manuals to support standardized data entry (e.g., hydrograph, taxonomy, etc.).
* Assess whether new or updated GIS features associated with these data are required (e.g., population boundaries, stream survey locations, fish data collection facilities, etc.).

In preparing data for submittal to StreamNet, the following quality assurance steps should also be implemented:

* Data checked against validation rules to meet requirements of the StreamNet and CAX DESs. Conversion of source data to the DES is the responsibility of the project’s data stewards that transfer data to the database. Adherence to the DES ensures data can be loaded into the database, queried accurately, and are equivalent for analysis by users (DES, <https://www.streamnet.org/resources/exchange-tools/>). Frequently this step is accomplished by making adjustments based on attempts to submit using the StreamNet API.
* Upon completion, data are preferably submitted using the StreamNet API.
* To preserve the quality control timestamp and StreamNet record ID, existing data records should be corrected/updated but not deleted from the data system. If assistance is needed, the StreamNet staff should be contacted.
* For metadata references that are reports or other publications, these documents should be sent to the Columbia River Basin Fish & Wildlife Library librarian (librarian) to establish a stable URL for inclusion in the StreamNet data record. If not done pre-submittal, StreamNet staff and the librarian will complete the task.
* For associated GIS features, review existing features available as web map services and displayed on PSMFC and StreamNet web tools for accuracy.

### Fields Requiring Special Attention

Identifying problematic fields during the QA phase can improve content by recognizing these fields require special attention. This is an ongoing process as the list of problematic fields will change over time. Some of these fields may be identified by the data steward managing their organization’s data while others may be revealed during independent visual quality control checks or by data consumers. To contribute to the overall quality assurance data stewards should highlight and communicate these problematic fields to the StreamNet and/or CAP Data Exchange Standards Development Team (DDT). This will contribute to overall data quality by making others aware that these fields may require more attention, and discussions about potential solutions to correct the problem can occur (e.g., clarifying or modifying the existing DES, and requesting assistance from the appropriate StreamNet committees, CAP Core Team, and/or the PNAMP Fish Monitoring Work Group).

Fields that are currently known to require particular attention, when preparing data for submittal to StreamNet, are listed in Appendix C. For these and any future problematic fields, the following steps should be employed to improve the data quality for such fields:

* Data stewards should highlight reoccurring problematic fields when requesting data from their organization's staff (see Appendix for example).
* To ensure proper completion, data stewards should provide guidance on what data is appropriate for these fields and any other information that can help ensure the fields are completed properly.
* These fields should be communicated to PSMFC-StreamNet for further discussion and resolutions.

## **Phase 2 - Quality Control**

Quality Control (QC), for the purpose of this plan, refers to the process of verifying that data records and data sets are complete, correct, and meet expected outcomes. Quality control focuses on identifying errors and omissions during and after submittal to StreamNet. Typical quality control tools include automated validation rules, post submittal spot-checks, and independent reviewers (modified from [2011 DAMA Dictionary](https://www.dama.org/cpages/body-of-knowledge)). During quality control, actions should at least include: 1) ensure fields comply with the validation rules, and 2) verify the submitted content is correct. Most of the quality control steps can be automated using routines, scripts, and queries that validate data based on specified quality requirements. However, human intervention is required to assess whether submitted content is appropriate for the field and if it can be easily understood by data consumers. An important element that contributes to a data consumer's level of confidence with data is documenting when QC checks are performed and when corrections occur.

### QC Procedures

The QC procedures include: 1) validation rules that are applied during the submittal of data to the StreamNet central data system, 2) visual checks that occur when data are in the StreamNet central data system, and 3) timestamps to communicate when QC checks are performed and updates occur.

#### Validation Rules

For records submitted to StreamNet, CAP Fish HLIs "CAX" and Fish Monitoring Data systems, automated validation routines are developed and maintained by StreamNet staff. Data are submitted one record at a time, with each record receiving multiple validation checks as follows:

**Level 1** - Each field has its own set of rules (i.e., ensuring numeric fields do not contain text, ensuring codes fall within the group of allowable values, and ensuring text strings are within acceptable length ranges).

**Level 2** - Ensure values in the various fields, within a data record, are compatible (e.g., if a record is submitted for spring run coho, it is rejected because there is no spring run for coho).

**Level 3** - Search for data problems, between rows of data within a table, to prevent duplicate data.

The interface used for data submission allows for adding new records, changing existing records, and deleting existing records. All validation failures result in rejection of the data record, as well as providing an error report with details explaining the rejection. Errors are conveyed back to the original data collectors, creating a feedback mechanism that promotes data quality improvements. A useful feature of automated validation routines is that data stewards can use it to validate the data before submitting to StreamNet (see [REST API Documentation webpage](https://www.streamnet.org/resources/exchange-tools/rest-api-documentation/)). This feature allows data providers to check, fix errors, and then submit a complete data set when it is known the records will pass validation.

#### Visual Checks

Visual checks of data in StreamNet's central data system complement the existing QC validation rules procedure. The visual checks focus on verifying whether content is appropriate and complete. This review is completed primarily by the data providers/stewards who submit data, other data stewards, and staff who serve as independent reviewers. To a lesser extent data consumers also perform visual checks when reviewing data for their use. The information reviewed, and input provided by these groups, reflects their level of familiarity with the content. For example, accuracy and/or validity of an organization’s data are best confirmed by the originating fisheries biologist/managers and associated data steward. Whereas independent reviewers are best suited to assess ease of comprehension and whether the non-biological data content matches what is intended by the DES. Also, data consumers who are familiar with a subset of the content may identify errors not detected by the other groups.

During the first implementation of the Visual Checks procedure, the below preliminary steps will be implemented to further refine this procedure:

* A "pilot review" will be conducted to identify the minimum number of records that each organization can process during a quarter and inform refinement of QC tasks. Eighteen records will be randomly selected and assigned to each data steward, representing about 3 records per HLI category from their organization's records that were submitted or updated between August 2020 to December 2021 records. The data stewards' tasks will focus on confirming accuracy of submitted content. Eighteen records will be assigned to an independent reviewer from the same subset of records assigned to the data stewards, but this review will focus on comprehension of content from a data consumer perspective, such as whether the comments and URLs are informative, and not on the accuracy of the biological data submitted. The time period selected aligns with records submitted since the July 15, 2020 DES version was adopted. Each individual will be asked to document the number of hours required to review each of the eighteen records and whether additional records could have been processed during the period. Through this accounting, it should be possible to estimate the number of records that an organization can review during a 12-month period based on the estimated time they can allocate to this task. This will also help inform task refinements and records selection, to improve efficiencies in reviewing records by the data stewards and independent reviewers.
* To identify what changes are needed to track the reviewed records, StreamNet staff will work with CAP DDT during the "pilot review" to determine an appropriate process to: 1) identify fields/records that need improvements and how best to convey that information with the submitting data steward, and 2) track and communicate, to data consumers, those records that have undergone the Visual Checks.

The following steps, which may be refined based on outcomes of the preliminary steps, are expected to be accomplished by data stewards and independent reviewers. We also describe how input received from data consumers will be addressed:

* Data Stewards:
  + When a data steward submits data, they should immediately review the new records to identify errors missed in previous steps.
  + Throughout the year, there should be a review of a subset of data entered in past years. This action will aid in identifying errors, as well as facilitate updates to previously submitted records. During this effort, records (recent and previous years’ data) will be compared to the original data in the organization's local data system. This task is essential since preliminary data, which is often provided initially, is eventually replaced with data that is characterized as final. The effort would include the review of records for completeness and accuracy (i.e., ensure the content of metadata fields correctly describe the submitted data and that the match between the metadata and data are sufficient to inform the data consumer, allowing for the proper use of the data).
  + The visual reviews can be performed manually or with the use of queries and scripts.
  + A visual check of the new GIS features for characterizing populations or groups of populations associated with new HLI estimates should be verified by viewing the “CAP Fish HLIs Query tool.” Checks of existing geometry should also be conducted to confirm that these are displayed as intended. Partners can also conduct this visual check by consulting the appropriate web map services directly.
  + A visual check of the associated map features describing the fish monitoring survey locations (trends) should be performed by consulting PSMFC’s published web map services or using the “Fish Monitoring Data Query tool”
* Independent Reviewers:
  + Independent reviewers, identified by PSMFC StreamNet in coordination with StreamNet members, will check a percentage of records annually to assess the appropriateness of submitted content, ease of comprehension, and whether information is missing based on field definitions. This review will not be focused on the actual biological data values but the supporting fields such as comment fields and metadata fields.
  + For those records that the reviewer identifies as problematic, the following highlights how issues will be communicated by PSMFC-StreamNet staff to the organization that provided the data.
    - StreamNet-funded organization - Correspondence sent to the StreamNet data steward that submitted the record.
    - CRITFC member tribes - Email sent to the CRITFC ITMD project coordinator with the expectation that they will work with the appropriate tribe.
    - BPA-funded organizations that do not have representation on StreamNet committees - Communication will be provided to the StreamNet COR to be forwarded to the appropriate BPA project COR.
    - Non-BPA funded data - StreamNet staff will contact the individual listed as the contact person for the record.
* StreamNet PSMFC staff
  + StreamNet staff will identify the records that will be reviewed during each quarter. To improve efficiency, records will be grouped based on similarities (e.g., a time series that uses the same MR.org protocol).
  + StreamNet staff will communicate with the appropriate tribal/agency data steward for issues identified by an independent reviewer.
* Data Consumers:
  + When a data consumer identifies an error, PSMFC-StreamNet staff will investigate the error to verify and recommend fields to be reviewed by the data provider, and as necessary, to correct the error.
  + If the data value is assessed by StreamNet staff as potentially requiring a correction, StreamNet staff will contact the appropriate data steward who will identify the underlying sources of the error and determine who (i.e., data steward or the biologist that provided the data) is responsible for making the correction.

#### Timestamps

Timestamps should continue to be used to denote the date and time for which records have been created and/or updated for individual data tables in both the CAP Fish HLI and the Fish Monitoring Data systems. All data tables should have the same set of timestamp related fields (i.e., DATASETVERSION, UpdDATE, SNLOADDATE, HLI\_LASTUPDATE, LASTUPDATED, and LASTMODIFIEDOn) that provide the following history for each record:

1. UpdDATE: When the record was last modified at the source organization. The data steward maintains this timestamp that they submit along with the data records.
2. SNLOADDATE: When the record was first added to central database.
3. HLI\_LASTUPDATE: Timestamp when an HLI was last updated in a specific record (only applies to CAP Fish HLI system).
4. LastUpdated: Most recent date/time when at least one DATA field was updated.
5. LastModifiedOn: Most recent date/time when at least one DATA OR METADATA field was updated.
6. DATASETVERSION: Version of the data set (Month, day, year, hour, minute.) This is a timestamp generated by the central database.

When issues are identified during the Visual Checks, the required corrections are completed at the source. Once resolved, the revised record is resubmitted by the data steward to StreamNet. This resubmittal results in an update to the timestamps #1, #3, #4, and #5 depending on the changes made to which fields. The Data Steward is responsible for updating timestamp #1 prior to the upload. Timestamp #2 should not change as it is a timestamp for the first time the record was added. To preserve the timestamp, such as timestamp #1 and other quality control related fields, existing records requiring correction are not to be deleted from the StreamNet central data system, but instead updated. As part of the Visual Checks pilot review, additional timestamps or other fields may be used to track and communicate when a record has undergone Visual Checks, regardless of whether a correction was made to the record and resubmitted.

### Fields Requiring Special Attention

During the Visual Checks the fields listed below should be verified as they tend to be associated with: 1) content that can be refined to better align with the intent of the field, 2) incorrect content, or 3) missing information. For these fields, the data steward submitting the content should confirm the accuracy and completeness, while the independent reviewer will assess the records for comprehension and agreement with the intent of the field, as described in the DES. Over time the fields requiring special attention will be updated as needed.

|  |  |
| --- | --- |
| **Field** | **Issue** |
| PopFit | *Incorrect use of designations* - “Same” is sometimes selected instead of “Portion”. The Same designation should only be selected when the estimate represents the entire population as defined by NOAA for ESA populations, or as agreed upon by the managers for non-ESA populations. Additionally, consider changing "Same" to "Whole" which may be a clearer term for most users. |
| PROTMETHNAME | *Relevant methods not provided* -Titles and name(s) of relevant methods are often not provided as defined in the DES. Instead, the information that is often provided in this field includes document citations for annual reports, project titles, etc., information that is appropriate for the PROTMETHDOCUMENTfield and not the PROTMETHNAME field.  Side note, the document associated with the PROTMETHNAME often does not include expected information such as study design, annual implementation notes if there were deviations from the study design, nor description of field methodology and analytical approaches. |
| INDICATORLOCATION | *URL functionality* - The URLs are often broken or do not allow users to access data due to being private. Also, some information available via URLs is not current. Yet, the URL is used as the source.  URLs are provided that take a user to a general website at which the user must “hunt” for the pertinent information.  URLs should be public and be the specific webpage with the relevant data. If this is not feasible, discussions are needed to explore what can be done to make this a useful field for the data consumers. |
| METRICLOCATION | *URL functionality* - The links are often outdated, incorrect, or broken. One of the major problems associated with the links is that they take the user to a general website that requires a query process or mining of listed reports. The links that are provided should redirect the user to the source of the data. |
| MEASURELOCATION | *URL functionality* - URLs take the user to a general website at which the user must “hunt” for the pertinent information. |

**Phase 3 - Quality Data Visualization**

The responsibility of ensuring the best available information is accessible to inform regional assessments, reports, and decision-making processes does not end with data QA/QC. StreamNet must also ensure data are communicated in a manner that best supports understanding and proper use by the data consumers.

### Effectively Communicating to Data Consumers

To provide quality information to data consumers, consideration must be given to how a data consumer may interpret these data. To ensure proper use, the data consumer must be able to understand these data. This understanding is influenced by what information is included and how it is organized and visualized whether on a map query, tabular query, or downloaded file. To ensure the content is communicated effectively, the following should be performed:

* Provide definitions (glossary) for all field names. A first iteration of this has been included the data download Excel file.
* For obscure or potentially misleading field names consider renaming or using aliases or labels that best describe the content in webtools and downloaded files, thus reducing the potential for a data consumer to make an erroneous assumption if the definitions are not consulted.
* Provide all metadata associated with the data set.
* When multiple fields and data sets are accessed at the same time (e.g., through a file download/export), data sets should be organized by data category (e.g., NOSA HLI).
* When multiple estimates are provided for the same fish group HLI, provide information that can inform the data consumer about how these estimates differ by populating a field that captures the difference in methods/analyses that are used for each estimate. This may require a new standardized field to complement existing methods 1, 2, etc. content.
* For fields that lack information, i.e., empty field, additional guidance needs to be provided so that the data consumer understands the empty field (e.g., missing data (age) due to time-lag in processing, or field not applicable for this record because data are not collected). For **fields that are not mandatory,** explore providing a code that indicates these data are not available or data are not generated. This action will reduce the potential to create the misperception of missing data that could lead to uncertainty and lack of trust in data quality/completeness.
* For content submitted at different times (e.g., NOSA submitted year x and age data submitted year x+1), fields should be separated to avoid confusion about data availability.
* Identify which fields should not be displayed or included in the download/export file due to being internal fields or obsolete fields (e.g. CBFWA name) that are of no value to the data consumer.

#### Fields Requiring Special Attention

For data consumers, uncertainty appears to exist relative to how the fields, listed in the following table, are displayed in the Excel file that can be downloaded from the website. How these fields are conveyed to data consumers should be assessed and refinements made.

|  |  |
| --- | --- |
| **Field** | **Issue** |
| PopFit | *Non-intuitive terms -* “Same” is not intuitive to data consumer as they are unclear what 'Same' refers to. The terms partial and multiple are more intuitive. |
| Lower Limit, Upper Limit, and Alpha | *Fields void of information*: For a data consumer, it is important to understand whether an empty field represents missing data. In the absence of an explanation, the data consumer expects that these fields should be populated with data. For those **fields that are not mandatory,** provide a code that indicates the data are not available or data are not generated |
| Including/Excluding Jack | *Data entry duplicated*: If it is unknown whether data are IJ or EJ, the data should not be entered in both fields. Clarity on how to populate this field for this scenario needs to be provided. |
| Age | *Field void of data*: Because there is a delay in processing samples for the purpose of aging, the age field cannot be populated concurrently with other year-specific data. Including age in the same table as other fields that can be populated sooner causes confusion due to the time lag for age and subsequent empty field. These fields should not be displayed on the same table or explanation should be provided to indicate age data exists and will be populated at a later time. |
| Various - URL fields | *URL functionality* - When a URL does not open a webpage that matches the expected content related to the field, data consumers can become confused and frustrated. |

## **Issues Beyond QA/QC Data Flow Tasks**

Data submitted to CAP Fish HLI StreamNet is an outcome of the collective work of numerous tribal, state, and federal biologist, managers, data stewards, and policymakers, reflecting agreed upon data to inform regional assessments and reporting. For some issues encountered by data consumers and occasionally data stewards, a need exists to include a broader group of representatives associated with the projects/programs that are sharing data with StreamNet. Depending on the specific issue, one of several existing forums could be leveraged for the needed discussion, including PNAMP's Fish Monitoring Workgroup (FMWG), StreamNet DDT, and CAP DDT.

The following is a summary of existing issues for which the identified groups could provide assistance:

|  |  |
| --- | --- |
| **Group** | **Issue** |
| CAP DDT and SN DDT | *Fields are not populated with the appropriate content or left void* - DDTs should collaborate to develop guidance for addressing the problematic fields to better communicate the desired content to the data providers (i.e., biologists/data stewards) |
| PNAMP FMWG in collaboration with StreamNet | *Fields are not populated with the appropriate content or left void -* a strong collaboration between biologists and data stewards, built on a common understanding of the correct content to be submitted for each field needs to be maintained over time. To this end, FMWG should work with StreamNet and PSMFC staff to convene a biennial technical workshop with biologists and data stewards to:   * Review DES fields (CA DES for HLIs and StreamNet DES for related trends) to ensure there is a common understanding of how to: 1) accommodate a "no methods" scenario to circumvent the required fields, 2) address fields with the wrong content, and 3) improve field descriptions that lead to confusion when entering data. * Discuss QA/QC issues and propose solutions. * Clarify expectations and roles relative to analyzing, providing, and submitting data. * Review fields that are frequently empty and identify causes and solutions (e.g., confirm data are not available or, if available, how to include the data in the submittal). * Receive input to clarify content and data provided through the query tools and the download/export file.   *Metadata and data set connections* - once automated metadata connections are made, review to ensure the connections are correct.  *Mismatch between StreamNet data fieldnames/values and original source* - explore how data consumers interpret existing mismatch and what can be done by StreamNet to provide needed content to alleviate confusion or distrust of data accessed from StreamNet. |
| PNAMP and StreamNet staff coordination | *Metadata and data set connections* - PNAMP and StreamNet staff should identify and implement more efficient approaches to connect metadata to data sets (e.g., project reports, MR title, and MR URLs). |
| CBF&W Library and StreamNet | *Access to data-related documents* - StreamNet staff, Library, and data stewards will ensure metadata documents are accessible to the public from the Library. The URLs provided by the data stewards should at a minimum have a library back-up URL if the data provider prefers maintaining their organization's document URL for public access.  *Process efficiency* - StreamNet and Library staff will implement changes to more efficiently link Library URL to datasets currently using other URLs or referring to printed documents. |
| NOAA, NPCC, BPA, and CAP Core Team | *CAP Fish HLI data in regional systems and reports* - assess whether the use of CAP Fish HLI data as a common data source across these products results in consistent information being displayed or whether there is a need to clarify / explain differences. |

## **Conclusion**

The procedures presented in this plan represent the minimum QA/QC that should be performed prior to, during, and after data are submitted to the StreamNet data system. Achieving and maintaining the highest level of QA/QC will require on-going vigilance and process refinement. Revisions are expected as new data providers and consumers become involved with the CAP Fish HLIs and Fish Monitoring Data. With the implementation of these procedures, data will be clear to users and their confidence in StreamNet as a source of fisheries data will increase..

Although the plan provides guidance to improve many aspects of existing QA/QC efforts, select issues are beyond the scope of this plan. Consequently, a comprehensive QA/QC effort will require assistance from StreamNet teams and from other groups who collaborate with the StreamNet Program

## **Appendix A: 2019 StreamNet Funded Data-Contributing Partners' QA/QC Procedures**

***Note: this content was created in 2019 and current processes may differ.***

Before data are shared with StreamNet, these first undergo QA/QC procedures at the source. The following section provides summaries of procedures used by current data providers who have StreamNet funded data stewards.

### Confederated Tribes of the Colville Reservation

The data that the Colville Tribes submit to StreamNet flows from collection in the field to StreamNet through the following steps: 1) field collection by Colville Tribes’ biologists, 2) data aggregation and calculations, 3) importing data to the Colville Tribe’s database, and 4) importing data to the StreamNet database. Quality assurance is ensured at the field-collection stage by using well established protocols, standardized training of field technicians, and the use of electronic data collection devices that use value constraints, pick lists, and required fields. All field data are collected by at least two crew members to ensure accuracy.

To ensure quality control, field data are reviewed daily and uploaded when appropriate, 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, while spreadsheets are checked for outliers by assessing value ranges or producing exploratory plots. Calculations that are embedded in data spreadsheets are checked for accuracy. Using a Python script, data are automatically uploaded to StreamNet.

### Idaho Department of Fish and Game

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

The QA/QC processes vary by data source and level of implementation. For collaborators, compliance with IDFG QA/QC processes is mostly voluntary. 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 QA/QC which occur at four stages between the collection of data in the field and its submittal to StreamNet. The QA/QC procedures occur at the user and programmatic levels 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 comparing local and central databases for completeness and correctness.

Programmatic-level QA/QC 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, and join definitions), orphan queries between parent and child tables, and duplicate queries within tables.

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

The data undergo further QA/QC 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 for completeness and correctness. 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 an efficient, linear data flow as the ETL and QA/QC processes include dynamic links between the data sources, local StreamNet SQL database, and regional SQL database.

Two exceptions to this data flow model are the Generalized Fish Distribution and the Columbia River Basin Fish Facilities, which are primarily used as geographic information system (GIS) data, and are compiled, managed, and shared as GIS data. The same QA/QC procedures apply to their tabular attribute data, but those are relatively limited and simple flat files. The geometric data (i.e., points, lines, and 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 StreamNet DES and submittal to StreamNet GIS, efficient and introduces less error. The GIS layers from IDFG and StreamNet are easily compared for completeness and correctness.

### Montana Fish, Wildlife & Parks

Montana Fish Wildlife & Parks (MFWP) implements 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 MFWP submittal to StreamNet. Quality assurance procedures occur at user and programmatic levels during data collection, entry, compilation, and submittal. User-level QA occurs during the collection of data and upon subsequent entry into the MFWP centralized database. Training and user guides have been created for field staff to facilitate consistent data entry. All field staff use standardized field survey forms when conducting netting, trawl, trapping, and electrofishing surveys. Data entered by the field staff into the agency’s centralized internal database are marked as provisional until these have been checked for accuracy. Once data have received biologist or supervisor QC, these are marked as reviewed. Only reviewed data can be analyzed 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 that are submitted to StreamNet undergo additional QA/QC. The data are transformed into a consistent format via a series of queries and then transferred to a secondary internal database. Agency StreamNet staff use a multi-part Python script to further QA/QC this dataset and ensure consistent procedures are used. The script performs a number of validation queries, compiles data into the necessary exchange table format, and then transforms data to meet StreamNet requirements. The data are georeferenced with the regional multi-scale hydrography, checked for logical consistency, orphans, duplicates, and then appended into a local StreamNet database. Upon completion, data are submitted to the StreamNet API where StreamNet performs real-time record-level validation before accepting data.

### Oregon Department of Fish and Wildlife

The Oregon Department of Fish and Wildlife (ODFW) utilize Oregon StreamNet staff within its Natural Resource Information Management Program (NRIMP) to request and gather reports, databases, and other data sets from around the state. The NRIMP is responsible for entering fisheries data into a SQL Server database via a web application that aligns with StreamNet data exchange standards, and utlizes validation rules developed for the StreamNet API. Data are obtained from 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, 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 given ODFW’s data management capabilities.

Oregon StreamNet staff apply QC processes during data entry and prior to submission to StreamNet. Updated records (between submissions) receive QC by comparing them to historic and current references. Data entry errors are corrected and records are marked for update, with the next submission. Data providers are notified of errors found in data they submitted, 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 visual 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 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 StreamNet DES, validation rules are applied, and checks are run on all required fields to maintain consistent and accurate information within the database. Staff use web application-based data entry screens, with instant validation of acceptable values via look-up tables. The ODFW database system also contains 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 that arise during data entry or QA/QC efforts, including calculation errors, appropriate use of DES codes, spatial accuracy, etc.

The submission process to StreamNet calls data exchange standards tables from the SQL Server database to be validated against the StreamNet SQL Server database. Several checks take place during this process including evaluations for duplicate primary keys, missing references for foreign keys, 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. Once 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 tool. During the submission process, ODFW’s data system automatically reads each submitted record back to validate it was transmitted correctly, and to compare record totals with StreamNet’s central data system to ensure the number of records matches.

### Washington Department of Fish and Wildlife

The Washington Department of Fish and Wildlife’s (WDFW) StreamNet data stewards have built and are building regional and corporate databases that store and report on collected data. These relational databases enforce referential integrity rules using drop-down pick lists, validation rules, and standardized input formats to control the information that is entered.

The WDFW StreamNet data mangers receive data from either corporate data sources like FishBooks (Hatchery Returns), Spawner Ground Survey, 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 are often provided in different formats. These include, but are not limited to, memos, electronic files, finalized agency reports, or personal communication. 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 available, to validate the original data.

When data meet the minimum requirements for the StreamNet DES, the data stewards enter the final data or transfer the data via electronic routines into the appropriate internal WDFW StreamNet database. Although the internal database includes fields that are useful for internal tracking and quality control, they 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 to validate the entries within prescribed limits and enforce referential integrity (i.e., rules established between tables). If data do not meet the minimum requirements of the StreamNet data exchange standards, the compiler contacts the source of the data to obtain complete information. Unfortunately, the biologist may be too busy to finalize the estimates. If the WDFW StreamNet data steward is equipped, they will apply expansion factors or run the raw data through known models to finalize an estimate and then return it to the biologist for a final approval. After 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 repeated for previous years data, as biologists may update data since preliminary numbers are occasionally submitted. Procedures enacted by each data partner must be carefully designed to ensure a smooth data flow exists from collection to the StreamNet data exchange. The procedures must include data validity checks to fill data gaps and correct erroneous values without introducing additional error through misinterpretation or accidental data manipulation.

When data are submitted to StreamNet, the compiler pulls a snapshot of the data from the database. A request is sent to the biologist with the snapshot of previously submitted data, and a set of instructions for issues of concern or new information that is required. They are asked to first to review previous years data and make any changes. Any issues that have occurred in the past are mentioned for fields that need attention such as a change in missing ages, methods, popfit, and to supply a current methods document for the new data if appropriate. If the biologists' schedule will not allow a timely turn around, then a date is requested when the new data can be expected. Note that many of the senior biologists responsible for this data are spending more time in the field due to reduced funding and staffing while still responsible for other duties. Lead times for the updated data are increasing from the time of the request.

The compiler checks the snapshot for completeness and accuracy while documenting the purpose of each table and any noteworthy issues. The data submitter further checks the data, making any final conversions that are warranted per StreamNet’s DES. References are sent to the StreamNet librarian. Identified data errors are sent back to the originator to correct the original data set.

## **Appendix B: Summary of 2021 interviews with Data Providers and Consumers**

### 

***Note: this content is based on an interviews performed in 2021. Some of the issues may have been resolved since completion of this review or were found to be non-issues.***

Multiple virtual interviews were conducted during April 2021. All interviewees were informed that the input received from these interviews would contribute to ensuring database fields and content meet the following expectations: 1) database fields are value-added, 2) available content is submitted to correct fields, 3); supporting content is provided and stable (i.e., URLs work), 5) FAIR principles are applied, 6) downloaded content is understood by data consumers, and 7) supporting content/metadata/URLs provide clarity. The focus of these interviews was on elements of the data and metadata that require a human to read and interpret (i.e., automated validation rules cannot prevent these issues).

### Data Consumers

Four data consumers that frequently visit the Coordinated Assessments Data Exchange (CAX) were interviewed during 2021 to learn about their experiences navigating through the CAX and accessing content.

The following questions were provided to all data consumers in advance of the interviews:

1. How have you accessed (viewed/downloaded) the data from the CAX?
2. Do you consult the (DES) when using data that you access/download?

What is your opinion of the End User License Agreement text? Does it help you understand what is expected if you choose to use these data?

1. How can we improve what is included with the data you access/download from the CAX to make it easier for you to understand and use these data?
2. Have you looked at the content under the field for the ProtocolMethods (name or URL)?
3. Have you attempted to use the metadata / URL provided to access the original data source on an agency/tribe’s database or report?
4. How can we improve the metadata so that you have the content you need to understand these data?
5. Are there some fields not necessary or missing?
6. Have you encountered anything else that led to confusion that we can improve? Any examples and suggestions for addressing?
7. Have you come across a population in CAX that had two or more different estimates?
8. We are developing a tabular query to complement the map query for CAX, are there some filters you would like to see included?

The following is a synthesis of the answers that were provided by the participants.

**1. How have you accessed (viewed/downloaded) the data from the CAX?**

Participants indicated that they have used the “download all data” function to acquire data from the CAX; however, the individuals indicated that they prefer to download the spreadsheets and perform their own filtering versus utilizing the CAX’s built-in query tool. One interviewee indicated that they download the data manually because they are unable to use R to perform the action automatically, unlike other online data sources that allow for such an action. The same individual suggested that it would be useful if “canned” or “template” JSON queries were available that would allow an inexperienced JSON user to access the data. The individual indicated that they have written scripts that allow for them to prepare the downloaded data; however, if StreamNet modifies the CAX in a manner that affects the structure of the downloaded content, the individual’s scripts will no longer function correctly.

**2. Do you consult the Data Exchange Standard (DES) when using data that you access/download? What is your opinion of the End User License Agreement (EULA) text? Does it help you understand what is expected if you choose to use these data?**

Results from the interviews demonstrate that not all data consumers review the separate DES document or EULA included in the data download Excel file. Some of the confusion around the field names highlights the importance of facilitating access to the content of the DES to the data consumers, or at least the definitions for the fields within the downloaded Excel file. Some interviewees indicated that the spreadsheets were initially difficult to comprehend. However, once the DES was reviewed, the user fully understood the meaning of the various fields. An interviewee indicated that the DES separates the CAP Fish HLIs (CAX) from other salmon data reporting sites, yet it could benefit from some modifications. For example, it was suggested that new data providers/consumers could benefit if there was a set of DES definitions for the layperson. For one interviewee, the field names are self-explanatory enough that a data dictionary is not warranted. Despite that opinion, the individual did suggest that there are specific-use definitions that are developed during DES meetings for which the subtleties (e.g., clarifying when age data includes HOF, NOF, or both) should be provided to the end users. Not all data consumers reviewed the DES and limited their focus to the content of the downloaded excel file or what is displayed on the query tool. Regarding the End User License Agreement (EULA), the interviewees provided responses that suggested that not all data consumers review the EULA.

**3. How can we improve what is included with the data you access/download from the CAX to make it easier for you to understand and use these data?**

The interviewees, regular CAX users, had strong opinions about the need for a function that allows the user to see what existing data have been updated, as well as new data that have been added to the CAX. Essentially, they suggested that some form of an indicator should be provided in the respective fields for which modifications/additions have occurred since the last system update. Also, an email notification and website highlight should be provided when a new DES is published. An effort should be made to thoroughly describe the updates.

As for the changes that allow for a better understanding of the data, an interviewee suggested that supporting information (i.e., local biologist knowledge) should accompany the data. For example, if a population record does not include any PHOS data after a given year, it would prove useful to the viewer if there was support information that explained the absence of PHOS, such as population did not receive HOF supplementation since date X.

**4. Have you looked at the content under the field for the ProtocolMethods (name or URL)?**

Responses from the interviewees indicated that they either have not or rarely click on the links. One interview revealed that they do not typically review the methods, but on the occasions that they do they have found that the metadata does not include all of the expected information. Although it is not common, multiple values for a single metric sometimes exist for a given year. One interviewee indicated that when they encounter such a situation, they use ProtocolMethods in an attempt to resolve the issue. The same individual indicated that they also utilize this field if there appears to be a change in methodology that could affect the manner in which data are used.

**5. Have you attempted to use the metadata / URL provided under the IndicatorLocation field to access the original data source on an agency/tribe’s database or report?**

This question resulted in varying responses. One individual stated that they never click on the links while others indicated that they have encountered links that are either broken or do not actually provide access to the data presented in the CAX.

The participants offered several approaches to consider that could help correct the broken link issue. They suggested that it should be the responsibility of the data provider to ensure the metadata are correct. Subsequently, it was proposed that CAX users should be able to contact the organization’s point person. Unfortunately, such an arrangement can be problematic due to the attrition of personnel. The interviewees all agreed that an ideal approach to prevent broken links is for the source documents to be stored in a “library.” Since all data sources will likely not be stored in a library, a participant suggested that there are approaches (e.g., scripts) that could be used to detect broken URLs and 404 issues.

**6. How can we improve the metadata so that you have the content you need to understand these data?**

Interviewees suggested that it would be useful if data consumers could submit comments directly regarding issues associated with specific records. Processes that will need to be considered include: 1) how comments are archived when new data are submitted, 2) access permission, and 3) reviews. A potential pilot effort could include select authorized users.

**7. Are there some fields not necessary or missing?**

This question resulted in an array of answers. Regarding fields that are not necessary, one interviewee indicated that there are fields that they never use. The reasons for not using the fields included inconsistent data or metrics that were not useful. The individual continued by stating that most fields could be useful if they were actually populated with data; however, since most data providers do not provide data, the usefulness of the field is lowered. Other participants suggested that disclosing why data are not available would be useful to the data consumer. Regarding missing fields, one interviewee suggested there is a need to add a field that defines point estimates. It is the individual’s belief that without a definition, the data consumer may assume the point estimates are maximum likelihood or the modes of a parameter’s probability density function. They suggested it would be ideal if a field existed that identifies a point estimate as “mode, median, or mean.” Other fields that received recommendations were harvest and broodstock removals. Participants suggested expanding these fields to better accommodate multiple fisheries/broodstock removal points and counts.

Although it was not part of the question, the interviewees were quite willing to identify the fields that are valuable to them which included: 1) Age, 2) Best Value, 3) Data type, 4) PopFit, and 5) Confidence Limits. One individual indicated that they appreciate the DataStatus field and suggested that the CAX may benefit from the addition of an additional field that informs the user that the data have been approved by the submitting agency.

**8. Have you encountered anything else that led to confusion that we can improve? Any examples and suggestions for addressing**?

The interviewees provided the following examples of issues that they have encountered while using the CAX, as well as some solutions:

* Abundance – The participants indicated that this metric (NOSA) is currently managed in a confusing manner. For example, NOSAIJ, NOSAEJ, TSAIJ, TSAEJ…are composites of information. Additionally, individuals referenced hatchery vs. natural spawners/escapement as being difficult to understand. One recommendation to correct the problem was to provide spawner abundance and then identify it as HOSA, NOSA, or Escapement.
* Origin – The interviewees found the multiple columns that reference jacks as problematic. They suggested providing the following fields so that the data are more tabular: 1) Abundance Estimate, 2) Origin, and 3) Including Jacks. They proposed that the table could be pivoted into the existing format.
* Estimate Type – This field, which defines whether estimates are “escapement” or “NOSA”, was characterized as nonsensical by the participants since it implies that if an estimate is spawner abundance (and not escapement), it is natural-origin. The interviewees suggested the descriptor should be “escapement” or “spawner abundance”. They indicated that the current descriptors have caused confusion among data providers.
* Escapement: “Above some Point” – Participants suggested that this is a really narrow way of defining escapement and that it does not work well in many situations (e.g., a NOAA-defined population that is a collection of small streams, each of which has a different mouth). With this scenario, the “point” is a collection of locations above a location at which escapement is measured.
* Best Value – Interviewees indicated that the Best Value field is useful but could be better defined. One interviewee suggested that it could be defined as “the best estimate for that year from the submitting agency, only including estimates where PopFit = same or multiple (partial are not best values).” They continued by stating that “it does not need to be the same method as other year’s best values and different submitting agencies can designate different values as best value.” The individual also suggested a need to ensure that data providers are not able to designate more than one best value per population per year, per origin, and with or without jacks. They finished by suggesting that each data provider should designate a “best value” for each year (if a data point meets the criteria).
* Downloads – Provide a data dictionary.

**9. Have you come across a population in CAX that had two or more different estimates?**

Some interviewees indicated that they have encountered the multiple estimate issue for many populations. One individual developed filters for the dataset, including a small number of “manual” hard-coded filters. The individual suggested that a better ruleset, for how data are uploaded into the CAX, may facilitate a more automated approach.

**10. We are developing a tabular query to complement the map query for CAX, are there some filters you would like to see included?**

Participants suggested that the easier it is to provide data set citation, the more likely that people will actually do it correctly. One individual expressed an interest in having the data set citation auto generated and that it would be included when data sets are downloaded. An additional comment that may improve proper use and citation of data sets was captured as follows:

*Citing data used that contains multiple years and multiple population data are challenging to cite. Simple data sets with one year/one pop is easier to cite. Generally, data consumer provides the URL from the data query and the accessed data as citation.*

### Data Stewards

A total of twelve data stewards/providers, representing three tribes and three states, were interviewed during 2021 to learn more about their QA process, discuss issues, and potential resolutions via QA and QC procedures.

The following questions were provided to all participants in advance of the interviews:

1. What challenges do you encounter when converting your organization’s data to conform with the DESs?
2. How can we reduce occurrence of outdated/incorrect documentation being linked to records such as the wrong data sources?
3. How can we reduce occurrence of links that do not redirect user to the specific web-content related to the record? Such as link to an organization’s homepage instead of a link to the page with the data-value or the report.
4. How do we handle situations that could lead to confusion from the data consumer’s perspective such as: when content in your system/source doesn’t match the content in the CAX (should we add your field name?); or the URL/source provided is to a private location they can’t access (should we hide from viewer? Or find a way to make that content accessible through the library or other way)?; field names (should we add your field name?)
5. Are there fields for which you don’t have data and don’t foresee ever having data?
6. Are there elements of the CAX records that you think can be vetted annually for QC by the StreamNet data stewards?
7. When you submit records to CAX (or StreamNet trends), do you resubmit whole data sets or submit only new/modified records?
8. To connect content across StreamNet and external systems we want to add a field for project numbers and annual contract numbers if relevant. Do you have project/contract numbers associated with data in your data system? If you don’t have project/contract numbers, do you believe that you would be able to add that information to your data system?
9. Have you downloaded CAX content or assisted someone else to download and interpret the content of the excel file?

The following is a synthesis of the answers that were provided by the participants.

1. **What challenges do you encounter when converting your organization’s data to conform to the DESs?**

The interviewees identified the following as data conversion challenges which are encountered when converting data to conform to the DES:

* Changes in lookup tables are not always communicated to the data providers.
* The DES is more basin-oriented, and this requires changes to the geography of the data.
* Methods are difficult to identify for trends data, especially the analyses. Thus, there is a

need to improve the linkage to methods.

* For juvenile traps, there could be a chance that runs cannot be separated (i.e., multiple

runs grouped together). Thus, there is a need for a mix-run option (if it does not currently exist).

* Changes to the DES “break” their programming. Subsequently, advance updates or   
  notifications would be useful. Also, it would be helpful if the DES/rules are more stable in the future.
* PopFit is a problem.
* People collecting data in the field do not participate in DES discussions.
* Since some indicators have several protocols, they are inserted in all of the fields.

An interesting and concerning challenge was identified related to the issue of obtaining standardized data. Some participants indicated that biologists do not always cooperate and provide/calculate data. As a result, data stewards are performing calculations. This likely explains why some of the values that are in CAX are not posted to an organization’s website. They continued by suggesting that the challenge is actually obtaining standardized data, not converting to DES. Because the biologists that are collecting the data are not involved in DES discussions, it is difficult to create a connection and understanding. To improve the data providers understanding of what is expected for a given field (e.g., PopFit), perhaps training would be beneficial.

1. **How can we reduce occurrence of outdated/incorrect documentation being linked to records such as the wrong data sources?**

Results from the interviews provided some concerning situations. For example, an interviewee indicated that they are not providing annual reports since they change every year. Continuing on that theme, another individual revealed that providing the metadata for a given value is the most challenging task for them as they do not always know where it is located. They indicated that it is easier for them to submit the data values. Also, an individual revealed that they do not provide links to the data because URLs change. As a result, they provide references that direct the user to data systems.

Understanding when data (e.g., age data) will become available is also important. Age data typically is not available when NOSA data are released. Subsequently, an interviewee suggested that providing a separate age data table would be beneficial since associated data are submitted after NOSA information is provided. They suggested this approach would allow for the data provider to avoid having to rewrite the NOSA records.

1. **How can we reduce occurrence of links that do not redirect user to the specific web-content related to the record?**

Answers from the interviewees provided some revealing information. First, two individuals indicated that they often are not aware of links that do not work and suggested that StreamNet staff should contact them when such an issue exists. Others suggested that the end users do not need the URLs for the source data, and that there are no references or documents for which links can be provided. For those that provide links, some indicated that they intentionally provide a general website URL instead of linking to specific pages since the links can often change or break. These individuals suggested that consideration should be given to adding a field that functions as a pathway (i.e., breadcrumbs) to the data source allowing users to navigate to the specific page where values are displayed. One potential approach that could provide clarification/assistance is that data providers could populate the “documentation” field and not the “URL” field while providing the report/location via breadcrumbs. A different approach, identified by one organization, was that they do not populate the ProtMethDocument field since the memo field is difficult to download. Consequently, they prefer to add a note informing the user to contact the data provider. Regarding the ability to access reports, some participants suggested that the ideal approach to avoid broken URLs would be for the reports to reside in an online library. Finally, it was suggested that a validation rule could be included that ensures the cells in the URL field contain an https or http.

**4. How do we handle situations that could lead to confusion from the data consumer’s perspective such as:**

* **when content in your system/source doesn’t match the content in the CAX**
* **URL/source provided is to a private location they can’t access**
* **field names**

Answers to this question were more of a “what could be done” response than what is actually being implemented to address the challenge. The following are examples that were provided by the interviewees:

* Add a field that allows for the inclusion of a short methods description explaining how data are collected. This addition may help the consumer better understand why there are two different values for the same population.
* Provide a spreadsheet that crosswalks agency and CAX field names.
* Internal agency links should not be in the ProtMethURL field. Instead, they should appear in one of the other ProtMeth fields (guidance currently exists in DES rules)
* The DES should indicate that private links are not to be assigned to the ProtMethURL field.
* For methods, use URLs that are associated with documents in a library.
* CompilerRecordID should not be part of download.
* Because multiple time series are problematic, a need exists to highlight associated records in a single time series (i.e., time series indicator for CAX data).
* Included with “data download”, there should be a tab that includes field names and definitions that are overviews of the DES (i.e., DES should not be repeated verbatim).
* Clarify the meaning of superpop/pop/MPG.
* Include NMFS’s popid.

1. **Are there fields for which you don’t have data and don’t foresee ever having data?**

This question resulted in conflicting responses from the interviewees. The following fields were identified as those for which data are not available and it is not expected that said data would become available:

* Confidence limits for trends that are a direct count (many of the studies were not designed to capture this metric)
* Confidence limits for jacks (includes hatchery fish), fractions (NOSA table), and age data (all tables)
* ProtMethDocumentation
* Age (these data are not available at the same time as NOSA)

One interviewee believes that all of the fields are important, regardless of whether they include data, since they function as breadcrumbs that help a user better understand the data. As a result, their organization attempts to populate as many fields as possible. The individual also suggested that it could potentially be useful if the data providers were informed as to why the fields are important for the data consumers. It was their belief that such information could lead to more fields being populated.

Contrary to the belief that all fields are important, several interviewees identified fields that should be updated or potentially removed from the CAX including:

* All comments fields should to be modified so text is not truncated (Comment text fields do not work in a spreadsheet).
* Inclusion of R/S juvenile and adult in the same table creates confusion due to orders of magnitude. It was suggested that this problem could be alleviated by assigning these values to different tabs.
* Remove the TRTmethod field.
* The BestValue field is difficult to understand, specifically when multiple agencies provide “BestValue=Yes” for a value for the same year. For the end user, confusion arises due to there being more than one best value.
* Remove ID fields (e.g., RunID).
* Remove ProtMethDocumentation fields.
* One interviewee suggested that up to 25 fields could be removed.

Throughout the interviews, there was a call for a data dictionary. Another suggestion was the addition of “pre-table paragraphs” that include field definitions.

**6. Are there elements of the CAX records that you think can be vetted annually for QC by the SN data stewards?**

The intent of the question was to determine what could be completed internally by the data stewards to improve QC. Unfortunately, responses from many of the interviewees suggest that the question was misunderstood. Several interviewees indicated that internal reviews, potentially by a different group of individuals, would be more productive than having the effort conducted by an outside organization.

For those that understood the intent of the question, they indicated that their organizations have established QC processes. One interviewee indicated that whenever new data are entered into the CAX, the data for the previous five years is reviewed for accuracy. Another individual described their organization’s process which includes performing a gross check by summing all values in their data system to ensure there is a match with the information submitted to the CAX. Basically, CAX tables are compared to local tables to detect errors.

The interviews did provide some interesting information regarding staff participation. One interviewee indicated that data collected by former staff members can create problems since new biologists prefer to not be responsible for data that they did not collect. Also, there was the observation that some organizations identify one individual as the contact for all data whereas, another organization found that to be potentially problematic. The organization that was critical of the process indicated that their organization identifies the project biologist for each record entered into CAX. They also stated that it would be beneficial if clarification was provided regarding whether the direct contact should be identified or if a generic contact is sufficient.

Some participants suggested relying on the users to provide QC/QA. The process would include an interactive feature that allows the user to contact StreamNet when they encounter issues. Finally, there were participants that indicated that they liked how StreamNet X was used for QA/QC efforts and that perhaps a similar service could be used for the Fish HLI.

**7. When you submit records to CAX (or SN trends), do you resubmit whole data sets or submit only new/modified records?**

This question resulted in mixed answers from the interviewees. Some data providers indicated that they annually provide “bulk” updates that replace all previous data while other providers indicated that they only provide new data for the most recent year. It appears that some of the data providers are planning to initiate the action of providing only new data. For one data provider that updates entire data sets, they indicated that by comparing data that is in the CAX with the entire data set that is annually submitted, they are essentially engaged in quality control. A common theme among the interviewees was the need for a timestamp. The timestamp, coupled with an audit field labeled “record visually QA”, would provide end users with confidence that the content in the CAX is correct regardless of whether the values and field names match the original data source.

**8. To connect content across StreamNet and external systems we want to add a field for project numbers and annual contract numbers if relevant. Do you have project/contract numbers associated with data in your data system? If you don’t have project/contract numbers, do you believe that you would be able to add that information to your data system?**

Most interviewees indicated that they provide project numbers; however, there were a couple of individuals that said they do not receive project numbers from the biologists or that the projects are not BPA-funded. To provide flexibility, individuals suggested that “funding source” and other “project”fields could be included to accommodate non-BPA projects

Regarding contract numbers, interviewees suggested that they could look into providing the information, especially if the effort could be used to automate a connection to the associated information. They did insist that there could be problems since single BPA-funded projects are often associated with multiple contracts.

With regards to using data presented in annual reports, one interviewee warned that the quality of data presented in annual reports can vary greatly. Consequently, caution should be used when connecting to annual reports.

**9. Have you downloaded CAX content or assisted someone else to download and interpret the content of the excel file?**

Responses from the interviewees indicated that some have and some have not provided assistance downloading/interpreting CAX data.

## **Appendix C: Manual Review of CAP Fish HLIs Data to Discover Issues, Assess, and Propose Solutions**

***Note: this content is based on an independent review of records performed in 2021.***

***Some of the issues may have been resolved since completion of this review***

***or were found to be non-issues.***

The dynamic nature of records submitted to StreamNet’s CAX (CAX) and the increasing number of records validates the need for a quality control (QC) process that includes the visual review of a percentage of records each year to complement the existing automated data validation processes. It is expected that through the visual review of the CAP Fish HLIs data, problematic fields will be identified and why these issues exist will be better understood. This information is likely to lead to improved QA and QC processes for submitting data and for conveying this information to the data consumers.

### Approach

An exploratory review of the CAX records was performed for BPA populations of interests, as well as additional records for species, runs, and populations confirmed by the PSMFC’s StreamNet program manager and fishery biologist. The objective of the review was to identify whether there are problematic fields associated with NOSA and OutJuv HLI records in the CAX.

Users have expressed concerns regarding select data fields (i.e., Age, Lower Limit, Upper Limit, and Alpha) that are often void of data. To identify the factors that are contributing to fields being void of information, 33 location/species/run queries were performed. Following the queries, data providers from Washington, Idaho, Montana, and Oregon were interviewed to assess the query results and better understand the factors contributing to fields lacking data.

### Records Reviewed

Table 1 - The BPA populations of interest (gray cells) and additional records for species, runs, and populations identified by the PSMFC’s StreamNet program manager and fishery biologist. The review was limited to data associated with NOSA and OutJuv HLIs:

|  |  |  |  |
| --- | --- | --- | --- |
| **Location** | **Species** | **Run** | **Entities Providing Data** |
| Entiat River | Chinook | Spring | WDFW |
| Kalama River | Chinook | Spring | WDFW |
| Clackamas River | Chinook | Spring | ODFW |
| Catherine Creek | Chinook | Spring | ODFW, NPT |
| Tucannon River | Chinook | Spring | WDFW, NPT |
| Salmon River Upper Mainstem | Chinook | Spring | IDFG, FPC, NPT |
| Lemhi River | Chinook | Spring | IDFG, NPT, FPC, BIOMARK |
| Coweeman River | Chinook | Fall | WDFW |
| Grays and Chinook Rivers | Chinook | Fall | WDFW |
| Wenatchee River | Chinook | Summer | WDFW |
| Okanogan River | Sockeye |  | FPC |
| Redfish Lake | Sockeye |  | IDFG |
| Lower Gorge Tributaries | Chum | Fall | WDFW |
| Grays and Chinook Rivers | Chum | Fall | WDFW |
| Upper Cowlitz River | Coho | Early and Late | WDFW |
| Clackamas River | Coho | Early and Late | ODFW |
| Lower Gorge Tributaries | Coho | Late | ODFW |
| Coweeman River | Coho | Late | WDFW |
| Grays and Chinook Rivers | Coho | Late | WDFW |
| Scappoose Creek | Coho | Late | ODFW |
| Okanogan River | Steelhead | Summer | CCT |
| Entiat River | Steelhead | Summer | WDFW |
| Wind River | Steelhead | Summer | WDFW |
| Hood River | Steelhead | Summer | ODFW |
| John Day River Lower Mainstem Tributaries | Steelhead | Summer | ODFW |
| John Day River Upper Mainstem Tributaries | Steelhead | Summer | ODFW |
| Umatilla River | Steelhead | Summer | ODFW |
| Tucannon River | Steelhead | Summer | WDFW |
| Grande Ronde Upper Mainstem | Steelhead | Summer | FPC |
| Lemhi River | Steelhead | Summer | IDFG, NPT, FPC, BIOMARK |
| Salmon River Upper Mainstem | Steelhead | Summer | IDFG, FPC, NPT |
| Kalama River | Steelhead | Winter | WDFW |
| Hood River | Steelhead | Winter | ODFW |
| Fifteenmile Creek | Steelhead | Winter | ODFW |

### Summary of Issues Discovered

The following issues were identified relative to the NOSA and OutJuv HLI records:

* Broken links to data sources
* Outdated sources
* Incorrect sources
* Data presented in CAX does not exist in the referenced sources
* Links that redirected the user to a general homepage instead of a specific location or report on the website
* CAX metrics do not align with those identified in the sources
* Multiple columns are consistently void of information
* Data often could not be verified

The occurrence of these issues was not the exception but rather the norm. Although the objective of the review was not to perform a review of the CAX data, random checks for data accuracy occasionally revealed differences between what was displayed in the source location and CAX.

### Problematic Fields

The following section provides an overview of the issues that were observed for the NOSA and OutJuv HLI records. The focus of the manual QC review was the data fields and, to a lesser extent the data itself.

For most data records, the fields identified in Table 2 were void of data. Through interviews with the data providers, it was learned that the monitoring designs for many of the projects were not developed to provide the requested data thus, the information will not be available in the near future. Also, data for age is not immediately available. Since there can be a 1+ year delay in the submittal of age data, a new “tab” should be considered allowing for the information to be viewed as a stand-alone data set.

Table 2 - CAX fields often void of information.

|  |  |
| --- | --- |
| NOSAIJLOWERLIMIT | HOSJF |
| NOSAIJUPPERLIMIT | TSAIJLOWERLIMIT |
| NOSAIJALPHA | TSAIJUPPERLIMIT |
| NOSAEJLOWERLIMIT | TSAIJALPHA |
| NOSAEJUPPERLIMIT | TSAEJLOWERLIMIT |
| NOSAEJALPHA | TSAEJUPPERLIMIT |
| NOBROODSTOCKREMOVED | TSAEJALPHA |
| PHOSIJLOWERLIMIT | AGE9PROP |
| PHOSIJUPPERLIMIT | AGE9PROPLOWERLIMIT |
| PHOSIJALPHA | AGE9PROPUPPERLIMIT |
| PHOSEJLOWERLIMIT | AGE10PROP |
| PHOSEJUPPERLIMIT | AGE10PROPLOWERLIMIT |
| PHOSEJALPHA | AGE10PROPUPPERLIMIT |
| NOSJFLOWERLIMIT | AGE11PLUSPROP |
| NOSJFUPPERLIMIT | AGE11PLUSPROPLOWERLIMIT |
| NOSJFALPHA | AGE11PLUPROPUPPERLIMIT |

**NOSA**

*ESTIMATETYPE* – The metric presented in this field often did not align with what was identified in the source documents (e.g., NOSA was listed in the field; however, escapement was the metric that was provided in the source).

*NOSAIJ, NOSAIJLOWERLIMIT, NOSAIJUPPERLIMIT, NOSAIJALPHA, NOSAEJ, NOSAEJLOWERLIMIT, NOSAEJUPPERLIMIT, NOSAEJALPHA* – Often, the values in the NOSA fields were described as escapement in the source document and not abundance. Also, rarely did the data sources (if provided) indicate whether the values included or excluded jacks as well as provide CI and alpha values. Regarding NOSA data, when data are provided by multiple organizations during the same year, for a specific population, occasionally the values are not the same. Better metadata documentation or improved communication of existing information should be applied to clarify why the values are different.

*NOBROODSTOCKREMOVED* – For most records, if this field contained data it was not available via the sources.

*PHOSIJ, PHOSIJLOWERLIMIT, PHOSIJUPPERLIMIT, PHOSIJALPHA, PHOSEJ, PHOSEJLOWERLIMIT, PHOSEJUPPERLIMIT, PHOSEJALPHA* – When data are presented, it is not available for viewing via the provided sources. The PHOS values typically are not provided in the sources. Instead, they appear to be manually computed using the NOSA and TSA values. Also, the lack of references to jacks in the source documents causes the user to be unsure of where the data originated and whether to trust its accuracy.

*TSAIJ, TSAIJLOWERLIMIT, TSAIJUPPERLIMIT, TSAIJALPHA, TSAEJ, TSAEJLOWERLIMIT, TSAEJUPPERLIMIT, TSAEJALPHA* - Rarely did the data sources (if provided) indicate whether the values included or excluded jacks as well as provide CI and alpha values.

*AGE* (all columns) – For many of the records, the source data (if provided) either did not provide the data that is displayed in CAX or lacked the CI and alpha values that are presented.

*INDICATORLOCATION, METRICLOCATION, MEASURELOCATION* – When provided, the links are often in the wrong column (i.e., PROTMETHURL, PROTMETHDOCUMENT), outdated, incorrect, or broken. One of the major problems associated with the links is that they take the user to a general website that requires a query process or mining of listed reports. The links that are provided should redirect the user directly to the source of the data.

**OutJuv**

*PROTMETHURL, PROTMETHDOCUMENT, INDICATORLOCATION* - When provided, URLs are often in the wrong columns. For example, the *INDICATORLOCATION* column is often void of links; however, what appear to be appropriate URLs for the *INDICATORLOCATION* column are often present in the *PROTMETHURL* or *PROTMETHDOCUMENT* columns. Too often, the links in the *PROTMETHURL* columnprovide no information relative to methods.

Many of the CAX records are products of BPA-funded M&E projects. For the BPA-funded data, at the least, identify the URLs for the annual reports located in cbfish, instead of referencing desktop "workbook".

One of the major problems associated with the links is that they take the user to a general website that requires a query process or mining of listed reports. The links that are provided should redirect the user directly to the source of the data.

*TOTALNATURALLOWERLIMIT, TOTALNATURALUPPERLIMIT, TOTALNATURALALPHA* – Data was occasionally provided for these fields. However, the sources that are listed in CAX do not include the data.

### Recommended Solutions and Existing Group

Proposed recommendations are the result of a review of a subset of CAX NOSA and JuvOut HLI records. The review resulted in the identification of a set of fields for which potential issues exist (i.e., data missing or not correctly assigned to the appropriate fields). These observations were vetted and discussed with data providers and consumers through interviews to gain insight on the nature of the issue and possible solutions.

From the interviews, it was determined that several of the issues raised by data consumers were not QA/QC issues since appropriate data do not exist. For many monitoring efforts, existing study designs do not provide opportunities to capture such information. Interviewees indicated that modifying said study designs to meet CAX expectations was not possible, especially for those projects not funded by BPA. Also, data providers indicated that for efforts that do collect samples for aging, missing age data are to be expected due to the amount of time required to process the samples. Consequently, the availability of age data may lag up to a couple years from when the samples were collected. Thus, missing lower/upper limits, alpha, and age data should not be characterized as QA/QC issues but instead a communication oversight.

Although the empty data fields did not represent a QA/QC issue, QA/QC problems were identified while reviewing the metadata fields (Appendix A). Despite the metadata fields being populated, the URL-centric fields (i.e., ProtMeth and Location) often do not contain links that function or effectively inform the user. While validation rules exist, they cannot effectively screen URLs. Thus, actions are needed to ensure the best information is entered into the StreamNet CAX. Appendix A lists URL-specific observations from the review and includes potential actions that would allow for a better evaluation of information sources and their subsequent inclusion in the CAX as well as the Columbia Basin Fish & Wildlife Library (previously named StreamNet Library). All referenced URLs and documents should be maintained in the StreamNet central data system and Columbia Basin Fish & Wildlife Library. Maintaining copies in the library will ensure users can access the documents regardless of a URL’s status.

**ProtMet**

|  |  |  |  |
| --- | --- | --- | --- |
| ProtMethName | The name(s) of all protocols and associated data collection and data analysis methods used to calculate the indicator estimate. | *Memo* | Provide title of protocol and name(s) of relevant methods used.  Documentation should describe the study design (including spatial, temporal, response and inference designs), annual implementation notes on variations from routine step by step procedures or design criteria, also known as survey design, description of field methodology and analytical approach. |
| *ProtMethURL* | URL(s) for published protocols and methods describing the methodology and documenting the derivation of the indicator. If published in MonitoringMethods.org, this link will provide access to study design information and all methods associated with the protocol. | *Memo* | Required if ProtMethDocumentation is null.  Provide URL(s) to source documentation of methodology. For MonitoringMethods.org provide link to the protocol. Methods documentation should include survey design, description of field methodology and analytical approach. URL links may be to online methods documentation resources like MonitoringMethods.org, other online resources, or online literature.  If methodology is unchanged from a previous year, use the previous link references. If methodology changed for this estimate, provide a new link. |
| *ProtMethDocumentation* | Citation or documentation that describes the protocol and/or method(s) listed in the ProtMethName field. Include references not documented in MonitoringMethods.org, such as reports, journal articles or other publications that describe the survey design, field methodology and analytical approach used to derive the indicator estimate. | *Memo* | Required if ProtMethURL is null.  Provide a citation(s) to documentation of the methodology used. This may be in the form of reports, journal articles, or other publications that describe the study design (including spatial, temporal, response and inference designs), variations from routine step by step procedures or design criteria, description of field methodology and analytical approach. If the methodology is not yet published, either insert here, or describe in a separate document and make it available online (provide the URL). Leave this field blank if methodology is described in MonitoringMethods.org.  Note: If there is no link to a cited document online, provide a copy of the document to the Columbia Basin Fish & Wildlife Library (streamnetlibrary.org). The library will scan the document and provide a URL. Post the URL in the ProtMethURL field.  If methodology is unchanged from a previous year, use the previous link or reference citation. If methodology changed, provide a new link or reference citation. |

**Observations and Potential Actions**

*ProtMethName*

Observation: Titles and name(s) of relevant methods are often not provided. The documentation does not describe study design, annual implementation notes if there were deviations from the study design, nor description of field methodology and analytical approaches. Instead, the information that is often provided includes document citations for annual reports, project titles, etc., information that is appropriate for the *ProtMethDocumentation* field.

* *Examples of incorrect entries (these examples do not constitute all incorrect entries)* 
  + Simpson, Philip C., Robert E. Reagan, and Hilary A. Doulos. 2017. Hood River Production Program Monitoring and Evaluation, Annual Progress Report. U.S. Department of Energy, Bonneville Power Administration, Project # 1988-053-04. (NOTE - The URLs for this entry, that are provided in *ProtMetURL*, takes the user to monitoringmethods.org and cbfish. The annual report, available via cbfish, should be saved to the Columbia Basin Fish & Wildlife Library.
  + Monitor Population-specific VSP Metrics for wild Chinook salmon in Idaho (1991-073-00) (NOTE - The URL for this entry, that is provided in *ProtMetURL*, takes the user to monitoringmethods.org)
  + A Compendium of Viable Salmonid Population Abundance and Productivity Field and Analysis Methods for Natural Origin Adult Chinook Populations in the Snake River Spring/Summer-Run ESU of Northeast Oregon from 1949 to 2019 (NOTE - The URL for this entry, that is provided in *ProtMetURL*, takes the user to a website that provides access to the document provided in *ProtMethName* field. This represents an example of a report that should be submitted to the Columbia Basin Fish & Wildlife Library. By including the document in the library, one does not have to be concerned about the link breaking, a fear that was mentioned multiple times during the interview process.
* *Examples of a correct entry (these examples do not constitute all of the correct entries)*
  + “Index area redds counts expanded to total. Used fish per redd value of 2.2 adjusted for the proportion of age3 fish in the run (Meekin 1966) modified by LaVoy" (NOTE – The URL for this entry, that is provided in *ProtMetURL*, takes the user to a website that provides the same text but does not provide a copy of the published documents (i.e., Meekin 1966 and LaVoy (unknown date). This represents an example of a report that should be submitted the Columbia Basin Fish & Wildlife Library. By including the document in the library, one does not have to be concerned about the link breaking, a fear that was mentioned multiple times during the interview process.
  + “Natural-Origin Spawners: Spawning escapement based on total census redd counts. Used Wenatchee fish per redd data to expand redds to spawners. Hatchery-Origin Spawners: Spawning escapement based on total census redd counts. Used Wenatchee fish per redd data.” (NOTE – The URL for this entry, that is provided in *ProtMetURL*, takes the user to a website that provides the same text but does not provide a copy of the published documents (i.e., Meekin 1966 and LaVoy (unknown date)).

Potential Action: Before submitting data, ensure study design descriptions are provided and not document citations. Electronic copies of referenced documents should be submitted to the Columbia Basin Fish & Wildlife Library.

*ProtMethURL*

Observation: URLs are provided that take a user to a general website at which the user must search for the pertinent information.

* *Examples of URLs that do not provide immediate access to data (these examples do not constitute all problematic URLs)*
  + The URL <https://fortress.wa.gov/dfw/score/score/> takes the user to an agency website for which the landing page lacks a structure that easily guides the user to the desired data. The link that could be used is <https://fortress.wa.gov/dfw/score/score/species/population_details.jsp?stockId=6742> since it provides direct access to the data.
  + The URL <https://odfw.forestry.oregonstate.edu/spawn/reports.htm> takes the user to an agency website for which the landing page provides a list of hot-linked annual reports. The link that could be used is <https://odfw.forestry.oregonstate.edu/spawn/pdf%20files/reports/18STWAnnualReport.pdf> since it provides direct access to the report from which the data were mined. (NOTE – This represents an example of a report that should be submitted to the Columbia Basin Fish & Wildlife Library (previously named (StreamNet library. By including the document in the library, one does not have to be concerned about the link breaking, a fear that was mentioned multiple times during the interview process.)

Potential Action: Before submitting data, check the URL to ensure the user has direct access to the data. If the links provide access to reports, provide an electronic copy to the Columbia Basin Fish & Wildlife Library to ensure there is an archived copy in case the URL becomes non-functional.

Observation: Monitoring Methods URL associated with draft (not public) document

* *Example of a Monitoring Resources’ URL that does not provide access to methods (this example does not constitute all problematic URLs)*
  + The URL [*https://www.monitoringresources.org/Document/Protocol/Details/2187*](https://www.monitoringresources.org/Document/Protocol/Details/2187) takes the user to the draft source that is not accessible

Potential Action: Before submitting URL, ensure the user can view the methodology section. To address draft documents and the inability to make them available to the public, consideration should be given to the development of a new field that classifies the URL or document as public or private.

**Locations**

|  |  |  |  |
| --- | --- | --- | --- |
| IndicatorLocation | Where this indicator is maintained at the source. | Memo | If online, provide URL(s). |
| MetricLocation | Where the supporting metrics are maintained at the source. | Memo | If online, provide URL(s). |
| MeasureLocation | Where the measurements are maintained that were used for these calculations. | Memo | If online, provide URL(s). |

**Observations and Potential Actions**

*Indicator Location*

Observation: URLs are broken or do not allow users to access data due to being private.

* *Examples of URLs that do not provide access to data due to their operational status (these examples do not constitute all problematic URLs)*
  + www.npt-cdms.nezperce.org
  + http://www.colvilletribes.com/obmep\_publications.php

Potential Action: Prior to submitting data, check URLs to ensure they provide the user with access to the data.

Observation: Information available via URLs is not current. Yet, URL is used as the source.

* *Example of a URL that does not contain current data yet it was entered as the URL (this example does not constitute all problematic URLs)*
  + Although the NOAA site <https://www.webapps.nwfsc.noaa.gov/apex/f?p=261:6>: only provides data as recent as 2015, it has been identified as the indicator source for 2016 – 2020 data.

Potential Action: Prior to submitting data, check URLs to ensure the data presented on the site aligns with that being provided to StreamNet.

Observation: URLs are provided that take a user to a general website at which the user must “hunt” for the pertinent information.

* *Example of a URL that does not provide immediate access to data (this example does not constitute all problematic URLs)*
  + The URL <http://odfwrecoverytracker.org/> takes the user to an agency website that requires an extended query process to obtain the information. The link that could be used is <http://odfwrecoverytracker.org/explorer/species/Chinook/run/spring/esu/282/285/> since it provides direct access to data for Catherine Creek spring Chinook.

Potential Action: Before submitting data, check the URL to ensure the user has direct access to the data. If the links provide access to reports, provide an electronic copy to the Columbia Basin Fish & Wildlife Library to ensure there is an archived copy in case the URL becomes non-functional.

*Metric Location*

Observation: URLs are broken and do not allow users to access data.

* *Example of a URL that does not provide access to data (this example does not constitute all problematic URLs)*
  + www.npt-cdms.nezperce.org

Potential Action: Prior to submitting data, check URLs to ensure they provide the user with access to the data. If only private URLs are available then explore providing additional content so data consumers know what to do/expect from these URLs

*Measure Location*

Observation: URLs are provided that take a user to a general website at which the user must “hunt” for the pertinent information.

* *Examples of URLs that do not provide immediate access to data (these examples do not constitute all problematic URLs)*
  + The URL <https://fishandgame.idaho.gov/ifwis/portal/> provides access to the agency’s landing page for IFWIS; however, an account is required to access the information.
  + The entry “Redd counts are located at <http://snq.streamnet.org/>” provides access to Streamnet’s “Fish Abundance Estimates and Indexes at Local Scales” site that requires an extensive query process to view the data.
  + The URL <http://odfwrecoverytracker.org/> takes the user to an agency website that requires an extended query process to obtain the information. The link that could be used is <http://odfwrecoverytracker.org/explorer/species/Coho/run/default/esu/159/166/> since it provides direct access to data for Sandy River coho.

Potential Action: Before submitting data, check the URL to ensure the user has direct access to the data. If the links provide access to reports, provide an electronic copy to the Columbia Basin Fish & Wildlife Library to ensure there is an archived copy in case the URL becomes non-functional. If only generic URLs are available then explore providing additional content so data consumers know what to do/expect from these URLs

**PopFit**

|  |  |  |  |
| --- | --- | --- | --- |
| **PopFit** | Categorization of how well the geographic extent of the NOSA/escapement estimate corresponds to the geographic definition of the population. | ***Text 8*** | This value must be "Multiple" if PopID represents a superpopulation.  Acceptable values: [*Do not include comments in brackets.*]   * Same [*Estimate represents one entire population, the whole population, and nothing but the population.*] * Portion [*Estimate represents a portion of one population. (Describe in PopFitNotes field.)*] * Multiple [*Estimate is from more than one population. (Describe in PopFitNotes field.)*] |
| ***PopFitNotes*** | Text description of how well the NOSA/escapement value corresponds to the defined population, and why the data are not at the scale of a single population. | ***Memo*** | This field is required if the PopFit field is "Portion" or "Multiple".  If the PopFit field is "Portion" or "Multiple", describe the lack of correspondence between the defined population and the fish for which the NOSA/escapement estimate was made. Also state *why* this scale of data was used to represent the population instead of true population-scale data. (Examples: "Data not available at exact scale of this population."; "Data at this scale best represent the population.") |

**Observations and Potential Actions**

*PopFit*

Observation: “Same” is sometimes selected instead of “Portion”. The Same designation should only be selected when the estimate represents the entire population as defined by NOAA for ESA populations or as agreed upon by the managers for non-ESA populations.

* *Example of an incorrect entry (these examples do not constitute all incorrect entries)* 
  + A NOSA HLI estimate Lostine/Wallowa Rivers spring Chinook population is identified as PopFit=Same (aka whole) but is described as only representing the Lostine River spring Chinook population. Therefore, the NOSA HLI representing only the Lostine River spring Chinook population should be entered as PopFit=Partial. Only a NOSA HLI representing the Lostine/Wallowa Rivers spring chinook population should be entered as PopFit = Same. <https://cax.streamnet.org/?species=Chinook%20salmon&run=Spring&popid=11##>
* *Example of a correct entry (these examples do not constitute all of the correct entries)* 
  + [*https://cax.streamnet.org/?species=Chinook%20salmon&run=Spring&popid=517##*](https://cax.streamnet.org/?species=Chinook%20salmon&run=Spring&popid=517) *a NOSA HLI estimated is provided for the* Naches River spring Chinook salmon that represents the entire population and is thus correctly identified as PopFit=Same (aka whole).

Potential Action: During annual workshops, include this field in the review of what is expected. Also, suggest exploring how ESA populations can be more clearly communicated to clarify when to select PopFit represents the whole (PopFit=same) population or when the population represents multiple separate populations (PopfIt = Multiple). Values assigned to PopFit should be verified by the data stewards to ensure the HLI estimate represents the population as defined by NOAA or as defined by the fish manager for non-ESA populations.

## **Appendix D: Boilerplate Template for Data Stewards to Communicate to Biologists the Expected Content for Problematic Fields**

***Note: that this is an example and the content***

***will need to be vetted with the latest DES version prior to use***.

Purpose: Clarify the expected content for fields that have been confusing in previous years when the biologists are asked to provide content.

When required fields have been ignored in the past or subfields have been a source of confusion in the past, the terms that have been skipped over or have caused the confusion in the past, are \*’d.

Below are examples of problematic fields and terms and the additional explanation that has been provided to biologists, which can serve as a template for other fields/terms over-time:

1. Does the data include Jacks?

You will see fields that look like they are duplicated. These are the **\*IJ** and \***EJ**. These suffixes mean including Jacks or excluding Jacks as in Chinook and Coho. **When you report species that do not have a jack portion, please put your data in the IJ fields.** While it sounds counter intuitive, this is the format for the NOSA table

1. What fish are you recording in the NOSA table?

This table stores information concerning natural origin spawner abundance (NOSA). "Spawner abundance" refers to ***the number of fish that actually spawn***, not necessarily the total number of fish returning to a spawning area -- ***all pre-spawning mortality has already been accounted for*** in the numbers represented in this table.

1. ***\*PopFit***

Popfit is the categorization of how well the geographic extent of the NOSA estimate corresponds to the geographic definition of the population you are entering data for.

You have 3 possible choices:

Acceptable values: [Do not include comments in brackets.]

* **Same** [Estimate represents one entire population, the whole population, and nothing but the population.]
* **Portion** [Estimate represents a portion of one population. (Describe in PopFitNotes field.)]
* **Multiple** [Estimate is from more than one population. (Describe in PopFitNotes field.)]

1. ***\*PopFitNotes***

This must be filled out if PopFit is Portion or Multiple. Describe how well the natural origin **spawner** abundance value you are reporting corresponds to **the defined population**, and why the data are not at the scale of a single population.

1. ***\*EstimateType***

You have 2 possible choices:

* **NOSA**

For NOSA – the number of fish that actually spawned. **Prespawn morts (fish that returned but died before spawning) are NOT included in this number.** They are the number of returning fish that survive to the location indicated in the WaterBody field.

* **Escapemen**t

The number of returning fish that survive to the location indicated in the WaterBody field that are **More Than** estimates of natural spawning abundance, where natural spawning abundance is the number of fish that actually spawned.

1. ***\*TRTmethod***

Flag indicating whether the methods used to generate the values in this record are those currently used for Endangered Species Act status reviews.

Acceptable values: Yes or No

1. ***\*MethodNumber***

This field represents the method(s) used to calculate the values in the "NOSAIJ" or "NOSAEJ" sections.

If only one set of methods is used to calculate the values for all years for a population, enter "1" for all records. Even if methods changed, you can enter "1" for all records if there is always only one record per year for a population.

**If more than one set of methods is used to calculate final values over a range of years for a population, use this field to indicate which records are meant to go together.** For example, if method 1 was used to calculate values for 1960 through 1994, and method 2 was used to calculate values for 1980 through 2013, then there will be more than one record for the years 1980 through 1994. In such cases you would enter "1" for records that result from the 1960-1994 method, and "2" for records that result from the 1980-2013 method. Similarly, if 3 different methods are proposed in an area for the same years, then use "1" and "2" and "3" to indicate which records belong together. **This lets a data user know which records belong together.**

1. ***\*NOSAIJ (incl jacks) or NOSAEJ (excluding jacks) Field 19***

Provide **whole numbers only**, not decimal values. Species for which jacks are recognized are Chinook salmon, coho salmon, chum salmon (rarely), and winter steelhead (rarely). Note that you can include upper and lower confidence intervals for this value. Report number and origin of jacks in the NOSJF (Proportion of natural origin spawners that are jacks) and HOSJF (Proportion of hatchery origin spawners that are jacks)

1. ***TSAIJ or TSAEJ***

Total Spawning Abundance, with or without jacks. These are the estimated total number of fish contributing to spawning in a particular year. Includes both natural origin and hatchery origin returns, and adult and jack age classes. Provide whole numbers only, not decimal values. For populations for which "jacks" are not recognized, enter the TSA estimate in this field.

1. ***AGE Data – General Comments***

* The age distribution must be derived only from the natural origin fish of the specific population this record represents. The age information may represent the exact group of spawning fish indicated in the NOSAIJ field, or a somewhat different group of fish. For example, the ages may represent the population as the fish passed a dam on their way to the spawning areas. Whatever may be the case, ensure this information is included in the protocol and method documentation section.
* The proportion of natural origin fish that were age 2,3,4,5 etc., (brood year +2, 3, 4, 5 etc.)
* Values must be between 0 and 1. **Express with 3 digits** to the right of the decimal point.
* These age fields contain proportions by age **for the natural origin fish**.
* They are **NOT** the numbers of fish actually aged. They are **NOT** the expanded numbers for a population. See the example provided in the Age2Prop field.
* The values of the Age2Prop through Age11PlusProp fields MUST sum to 1.00 ± 0.01.

1. ***Documenting Methods Used to generate the Data – General Comments***

These are critical fields, and it is important that you complete at least for one of these. You have 3 different fields in order to clearly let the user of your data know how the data you reported were generated.

* In the field ProtMethName. A memo field so you are not limited by any number of characters. Provide title of protocol and name(s) of relevant methods used. Documentation should include the study design (including spatial, temporal, response and inference designs), annual implementation notes on variations from routine step by step procedures or design criteria, also known as survey design, description of field methodology and analytical approach.
* If posted at a website, then use the ProtMethURL field to point the user to the report or document that has the methodology in it.
* ProtMethDocumentation field is another memo field that has a Citation or documentation that describes the protocol and/or method(s) listed in the ProtMethName field. Include references not documented in MonitoringMethods.org, such as reports, journal articles or other publications that describe the survey design, field methodology and analytical approach used to derive the indicator estimate.
* The MethodAdjustments field is where you document “tweaks” or minor adjustments to the method described for the data documented in the time series Method field.

1. ***\*OtherDataSources***

If another federal agency or tribe created or conducted surveys in association with this data please fill in which tribe or agency participated. If you have co-managers this is where you put their name in.