

Protocol for Monitoring Springs at Ozark National Scenic Riverways, Missouri. Heartland I&M Network

SOP 11: Data Management

Version 2.0 (06/18/2019)

Revision History Log:

Previous version number	Revision date	Author	Changes made	Reason for change	New version number
1.0	6/18/2019	D. E. Bowles, H. R. Dodd, J. A. Hinsey	Changes made to reflect changes made in databases	Databases were revised and updated	2.0

This SOP explains procedures for data management of Ozark National Scenic Riverways springs community (fish, macroinvertebrate, and aquatic vegetation) sampling data. The springs community databases described in this SOP were created when the original springs database underwent a complete redesign that resulted in springs data being merged into the existing aquatics databases (fish and invertebrates) with a new database being created for aquatic vegetation as shown in Figure 1. This redesign was to reduce the number of aquatic databases requiring ongoing maintenance. This SOP includes a general description of the design, and procedures for data recording, data entry, data verification and validation, and data integrity for the primary invertebrate sampling data. More detailed step-by-step procedures on data entry will be kept on file in the HTLN aquatics lab located at Missouri State University.

Data management can be divided into (1) the initial design phase that involves defining the data model, its entities and their relations and (2) the procedures necessary to implement the database. Currently Microsoft (MS) Access is the primary software used for maintaining springs data. Water quality data will be stored in the National Park Service's water quality database (NPStoret, see Water Quality SOP 4). Environmental Systems Research Institute (ESRI) ArcGIS is used for managing spatial data associated with field sampling locations. Data products derived from this project will be available at the NPS IRMA Data Store and EPA water quality National Data Warehouse. QA/QC guidelines in this document are based on the NPS Inventory and Monitoring document *Certification Guidelines for Inventory and Monitoring Data Products* (NPS, 2016).

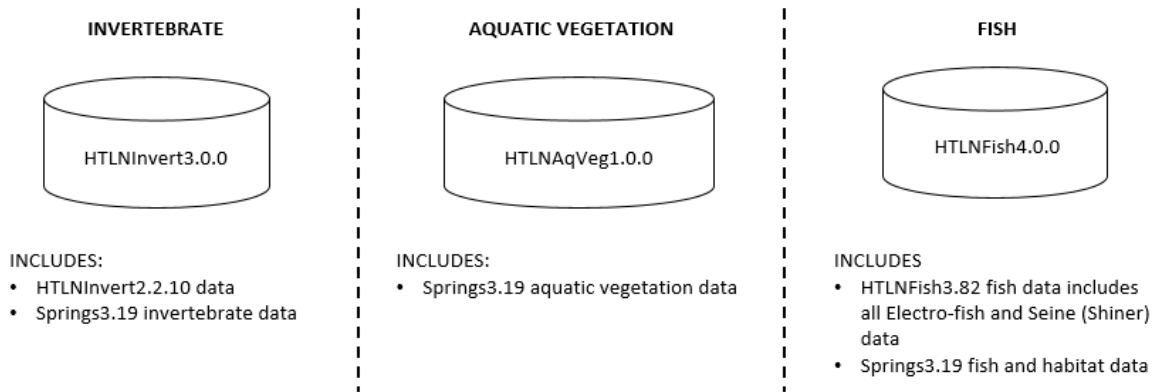


Figure 1. Illustrates how Springs3.19 data were merged into existing aquatic databases (fish and invertebrates) and a new Aquatic Vegetaion database was created.

Design

The data model and relational database systems for this project was created based on the requirements and standards for monitoring data systems at the NPS Inventory and Monitoring Division (Frakes et al. 2015).

Generally, the database model is developed from the appropriate field forms based on the relevant sampling protocols. Prior to table development, the data manager coordinated with project staff to determine data types (byte, text, numeric, for example), precision, and range of values. Acquiring field data forms assists with developing data logic (that is, how data components relate) and identifying indexes and primary keys.

A generalized model of the aquatic community databases includes two primary tables for sampling events and locations. These two core tables contain general information pertaining to the field sampling occasion (the when and where of the sample) and are present in all three aquatic databases. This includes information such as date and time, reach ID, and park/project codes. The aquatics community tables serve as the organizing hub for all data (Figures 2-6). Other tables primarily address habitat or water quality conditions. See Figures 2-6 and Table 1 for a cross reference of the field forms to the corresponding tables.

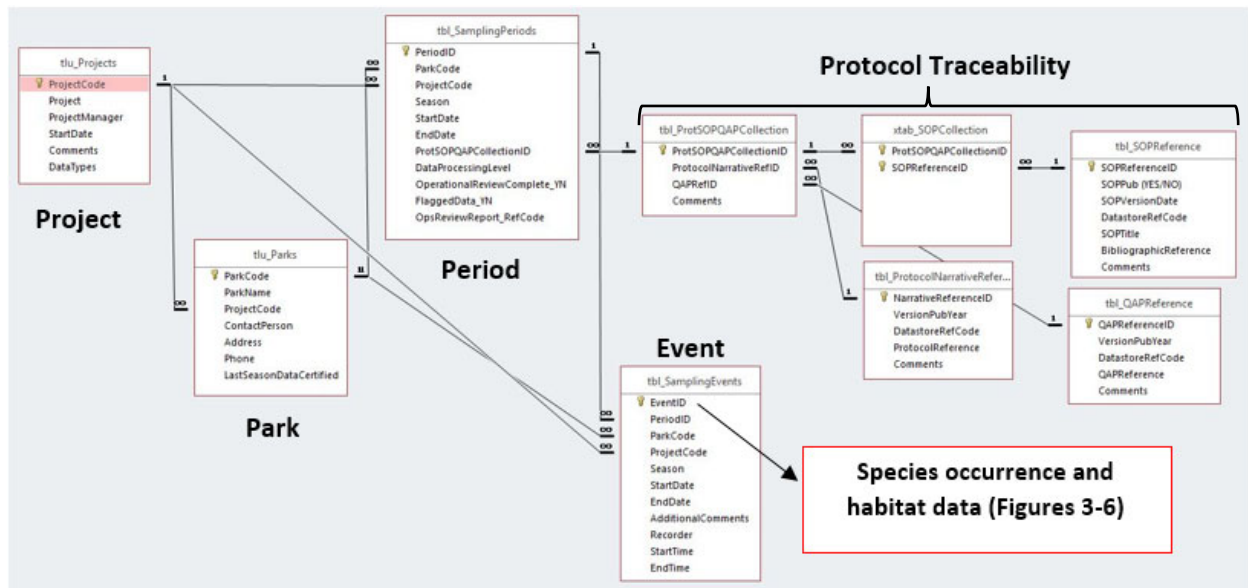


Figure 2. Entity relationship diagram of the HTLN aquatic common data tables present in all three aquatic databases (fish, invertebrate and aquatic vegetation).

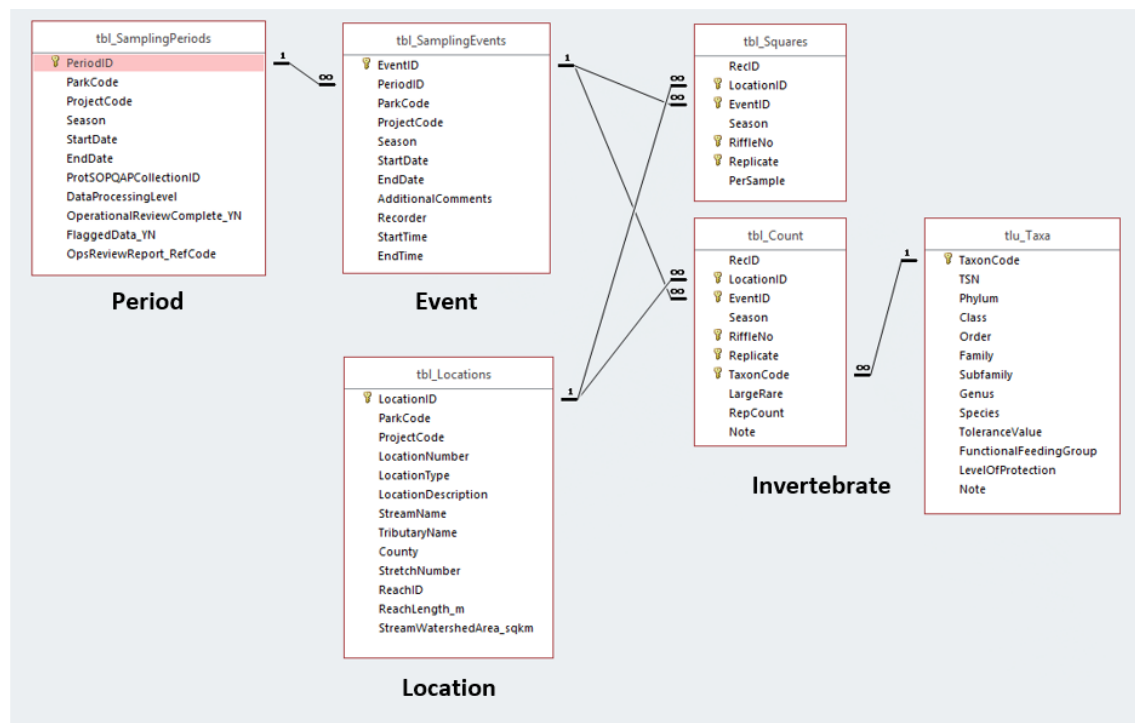
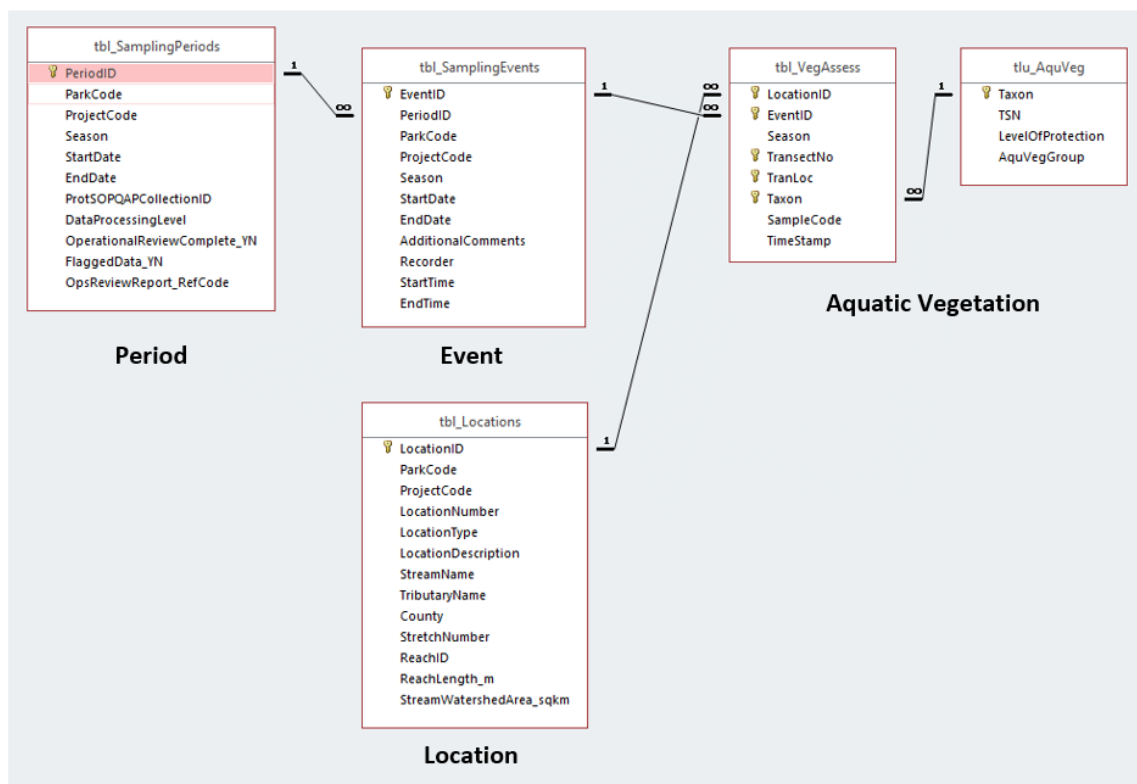


Figure 3. Entity relationship diagram of the HTLN invertebrate (HTLNInvert) database. Project data tables consist of species occurrence data (tbl_Count and tbl_Squares).



Figure

4. Entity relationship diagram of the HTLN aquatic vegetation (HTLNAqVeg) database. Project data tables consist of species occurrence data (tbl_VegAssess).

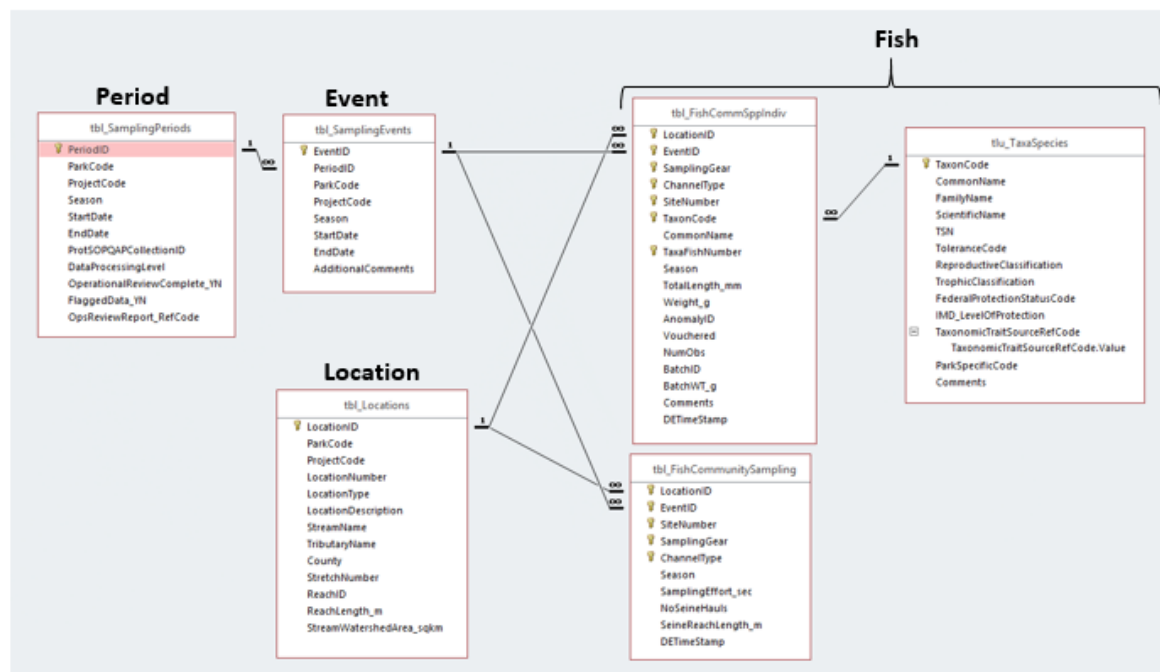


Figure 5. Entity relationship diagram of the HTLN fish (HTLNFish) database. Project data tables consist of species occurrence data (tbl_FishCommSpplndiv and tbl_FishCommunitySampling).

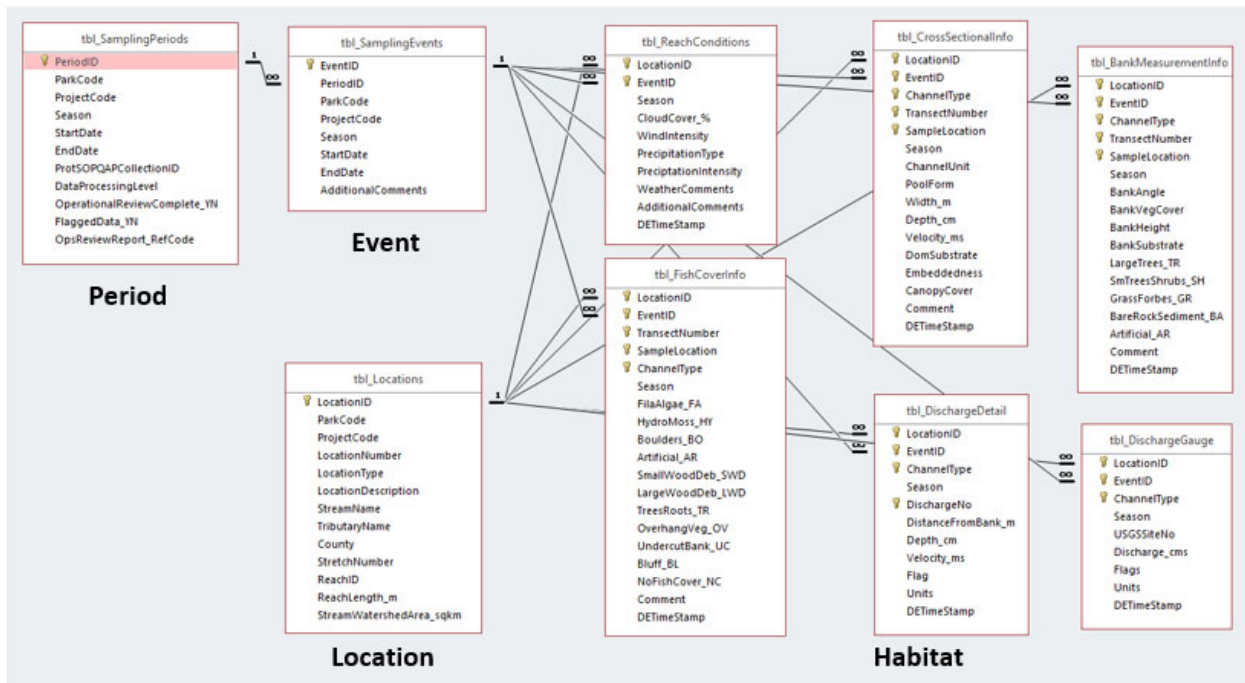


Figure 6. Entity relationship diagram of the HTLN fish (HTLNFish) database. Project data tables consist of habitat and other related data (tbl_ReachConditions, tbl_FishCoverInfo, tbl_CrossSectionalInfo, tbl_BankMeasurementInfo, tbl_DischargeDetail, and tbl_DischargeGauge).

Table 1. Cross reference table for aquatic database tables and corresponding springs field forms.

Data Type	Springs Field Form	Form Reference	Data Description	Database Name	Database Table Name
Protocol	n/a - determined based on the dates of all events during the sample season	n/a	Beginning and ending dates of the sample season	HTLNInvert, HTLNFish, and HTLNAqVeg	tbl_SamplingPeriods
Event	n/a - determined based on the first date of the sample event	n/a	A specific date a sample event occurred	HTLNInvert, HTLNFish, and HTLNAqVeg	tbl_SamplingEvents
Location	n/a - determined based on location within park boundaries	Narrative, Table 1	A specific location (site) of a sample event	HTLNInvert, HTLNFish, and HTLNAqVeg	tbl_Locations
Invertebrate	Aquatic Invertebrate Identification & Enumeration Sheets	SOP 7	Aquatic invertebrate identification count data	HTLNInvert	tbl_Squares, tbl_Count
Aquatic Vegetation	Aquatic Vegetation Assessment Form	SOP 5	Aquatic vegetation identification data	HTLNAqVeg	tbl_VegAssess
Fish	Fish Community Field Data Sheet	SOP 8	Channel type sampled, gear and effort, and individual fish length/weight data	HTLNFish	tbl_FishCommunitySampling, tbl_FishCommSpplIndiv
Habitat	Reach and Weather Conditions	SOP 9	Weather conditions	HTLNFish	tbl_ReachConditions
Habitat	Instream Habitat Assessment Form	SOP 9	Instream habitat data per transect	HTLNFish	tbl_CrossSectionalInfo
Habitat	Fish Cover Form	SOP 9	Fish cover types within the stream and along banks	HTLNFish	tbl_FishCoverInfo
Habitat	Bank Measurement Form	SOP 9	Bank stability and vegetation cover per transect	HTLNFish	tbl_BankMeasurmentInfo
Habitat	Discharge	SOP 9	Interval, depth, and velocity measurements or USGS gauge measurement	HTLNFish	tbl_DischargeDetail, tbl_DischargeGauge

Data Recording

Quality Assurance/Quality Control (QA/QC) procedures related to data recording, verification, validation and processing are important components of any project sampling data, including sample methods, effort, weather, water quality conditions, species abundance data, and habitat data that are recorded on the field form (see Springs Field Form and Form Reference columns in Table 1). QA are procedures that ensure data collection, products, and services have minimal errors. QC are procedures to ensure that data or products meet established data quality standards (see *Integrating QA/QC into the IMD Pipeline*, NPS, 2016). To ensure data is recorded properly on field sheets and to reduce data recording errors, consistency among the persons recording data is necessary and any new data recorder should be trained by the project leader on how to properly fill out the field sheets. During the training process, the project leader or HTLN staff that have been trained (Aquatic Ecologist) by the project leader in data recording should periodically check the quality of data recording by the new crewmember while the data is being collected at a site. Field forms are checked for completeness by the project leader or trained HTLN staff member after all data has been recorded and before leaving a site or within 24 hours of data recording. This will aid in verification and validation of the data after entry into the database. Verification is the collective processes that ensure data maintain accuracy and completeness throughout the data management cycle. Validation are the collective processes to evaluate legitimacy of the data throughout the data management cycle. Examples of validation procedures include: range checks based on environmental limits, comparison of data to previous data, and comparison of data to expected data or modeled expectations.

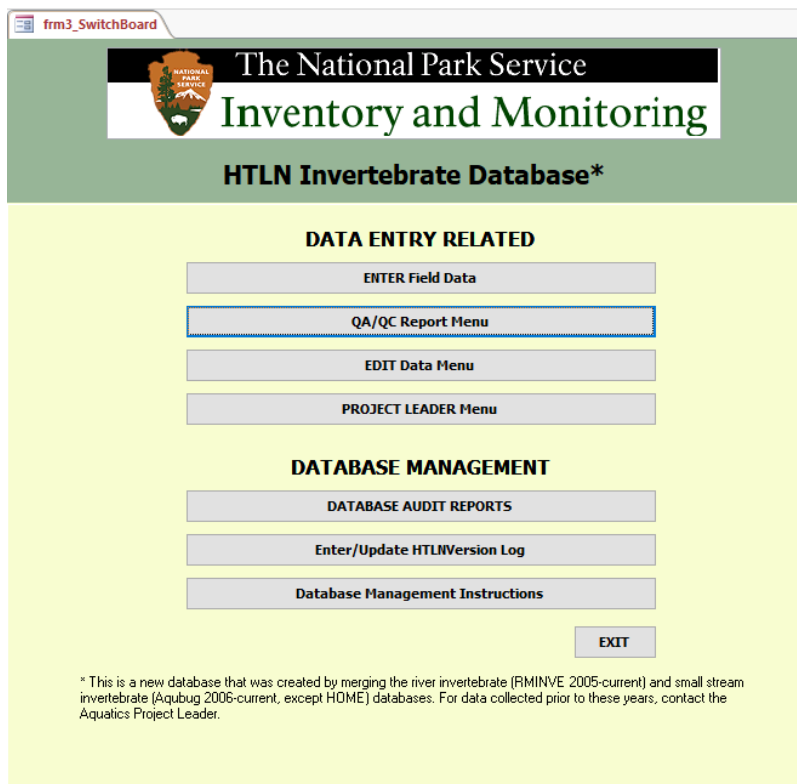
To prevent the complete loss of field form data because of unforeseen circumstances (for example, fire or flood in the workplace). Field sheets are scanned into a computer server that is located in a separate physical location from the original datasheets. A backup of the scanned field sheets is kept offsite, which ensures a copy of the field sheets are in at least three physical locations.

Data Entry

To reduce data entry errors and assist in the verification process, data entry will be completed by a staff member who collected the data or is familiar with the field collection process. A copy of the database is downloaded from the server on to the staff's harddrive for data entry, and this copy is renamed with the new version number. Upon opening the database, the user is presented with a switchboard that can be used to access forms for data entry (Figure 7). A preliminary set of forms define the sampling occasion and requires the input of location, period, and event IDs prior to entry of additional data. The other forms (Figures 8-10) address the details of species occurrence, habitat, etc. Once all fields for the preliminary set of forms have been completed, data can be entered for the remaining forms.

Several features are "built-in" to form properties that enable the user to maximize data entry efforts while minimizing error. These include: data input masks for ease of viewing multi-part data (such as park/project codes and date in PeriodID, LocationID, or EventID), "fill-in-as-you-type" to automatically complete a field, default values to autopopulate common values, limiting input values to known ranges (or restricting null values) or providing "drop-down boxes", and tab indexes to control the order of data entry. Forms also contain fields that require data input or are constrained to properties and integrity of related tables.

More detailed step-by-step procedures on data entry will be kept on file in the HTLN aquatics lab located at Missouri State University.



frm3_SwitchBoard

The National Park Service
Inventory and Monitoring
HTLN Invertebrate Database*

DATA ENTRY RELATED

ENTER Field Data

QA/QC Report Menu

EDIT Data Menu

PROJECT LEADER Menu

DATABASE MANAGEMENT

DATABASE AUDIT REPORTS

Enter/Update HTLNVersion Log

Database Management Instructions

EXIT

* This is a new database that was created by merging the river invertebrate (RMINVE 2005-current) and small stream invertebrate (Aqubug 2006-current, except HOME) databases. For data collected prior to these years, contact the Aquatics Project Leader.

Figure 7. Example of aquatic

invertebrate database field data entry menu.



frm1_EnterData

The National Park Service
Inventory and Monitoring
HTLN Invertebrate Database

ENTER Field Data

ENTER Core 4 and Riffle Data

ENTER Discharge Detail Data

ENTER Discharge Gage

ENTER Invertebrate Counts

Return to Previous Menu

Figure 8. Example of aquatic

invertebrate database count data entry menu.

The screenshot shows a web application interface for "The National Park Service Inventory and Monitoring HTLN Springs Aquatic Vegetation Database*". The interface is divided into two main sections: "DATA ENTRY RELATED" and "DATABASE MANAGEMENT".

DATA ENTRY RELATED

- ENTER Aquatic Vegetation Field Data (highlighted with a blue border)
- QA/QC Report Menu
- EDIT Data Menu
- PROJECT LEADER Menu

DATABASE MANAGEMENT

- DATABASE AUDIT REPORTS
- Enter/Update HTLNVersion Log
- Database Management Instructions

EXIT

* This is a new database that was created using the HTLNInvert3.0.0 database. Removed all tables, queries, forms, and reports not related to Springs Aquatic Vegetation, modified the design to accomodate Springs Aquatic Vegetation data, imported Springs Aquatic Vegetation data.

Figure 9. Example of aquatic vegetation database field data entry menu.

The screenshot shows a web application interface for "ElectroFish ENTER Fish Community Data". The interface is divided into three main sections: "1. SELECT Park Code, LocationID, and EventID:", "2. ENTER Fish Community Sampling Data (Sampling Effort, Seine Hauls and Seine Reach Length):", and "3. ENTER Fish Community Species Individual Data:". The "Park" dropdown is set to "main". The "LocationID" and "EventID" dropdowns are set to "main". The "Sampling Gear" dropdown is set to "backpack". The "Channel Type" dropdown is set to "main".

ElectroFish

ENTER Fish Community Data

1. SELECT Park Code, LocationID, and EventID:

Park: main

LocationID: main

EventID: main

Sampling Gear: backpack

Channel Type: main

2. ENTER Fish Community Sampling Data (Sampling Effort, Seine Hauls and Seine Reach Length):

ENTER Fish Community Sampling Data

QUERY Fish Community Sampling Data

3. ENTER Fish Community Species Individual Data:

ENTER Fish Community Species Individual Data

Close Form

Figure 10a. Example of fish database fish community and species individual data entry menu.

**ElectroFish
ENTER Reach Data**

1. Choose Park Code, LocationID, and EventID:

Park

LocationID

EventID

2. SELECT type of Reach Data to enter:

Weather Conditions

Core 4 WQ Discrete Measurements

Discharge Detail

Discharge Gauge

Reset Form

Close Form

Figure 10b. Example of fish database reach habitat data entry menu.

**ElectroFish
ENTER Transect Data**

1. Choose Park Code, LocationID, and EventID:

Park

LocationID

EventID

Channel Type

2. SELECT type of Reach Data to enter:

Transect Spacing Intervals

In-Stream Habitat

Bank Measurement Info

Fish Cover Information

Reset Form

Close Form

Figure 10c. Example of habitat transect database field data entry menu.

Entering Water-Quality Data

Unattended Core 5 water quality data collected by unattended data loggers or sondes (see SOP 4) are uploaded from the logger using the manufacturer's accompanying software program and stored on a the HTLN server. Data then are edited to correct for any missing data due to logger maintenance (downtime), corrected for drift since last calibration, and validated to determine if the data meet the expected range requirements or critical limits. Unattended CORE 5 water quality summary data then are entered into the NPS water quality database (NPSstoret). Coordinate data for logger locations are collected in accordance with the current HTLN spatial data collection techniques and stored in the NPS water quality database. Metadata are then entered for each characteristic/parameter. The water quality data is sent annually to the NPS WRD staff for initial QA/QC and subsequent upload into the WRD master copy of the EPA water quality National Data Warehouse database. See the Water Quality SOP 4 for details on data download, correction, and import.

Data Verification

Data verification immediately follows data entry and involves checking the accuracy of computerized records against the original source, usually paper field records. Data tables are queried to produce hard copy reports of specific sets of data (for example, habitat data). These reports then can be compared to the original field data sheets to identify missing, mismatched, or redundant records. One hundred percent verification between field sheets and the database tables will be conducted by a staff member familiar with data collection but who did not enter the data. Any corrections will be completed by the project leader. Following these procedures, 10% verification between field sheets and the database tables will be conducted by the project leader. If errors are present in the 10% verification processes, the data in question will undergo another round of 100% verification.

Data Validation

The purpose of data validation is to ensure the correct values of data are based on logical, expected or known prior values. Although data may be correctly transcribed from the original field forms, they may not be accurate or logical. For example, field crews collect data per occasion (location and event) and the resident data should reflect this. At any given occasion, sampling is conducted in each reach during the sampling event and can be validated by querying the ID number for these values. A query of these data should reflect these conditions and confirm the coincident (relational) nature of the database.

As data are amassed in the database, validation is conducted by query and comparison among years to identify gross differences. For example, species "A" may be recorded at a location in a specific year, but not in previous years, thus representing a possible new locality or potentially a mis-identification. In addition, range checks should be completed to ensure values on the field forms and in the database are within typical environmental limits. Validation should be conducted by the project leader, and any questionable values should be confirmed with data recorder and/or crewmember that collected the data.

Once verification and validation are complete, the dataset is turned over to the HTLN Data Manager for archiving and storage. These data then can be used for all subsequent data activities.

Spatial validation of sample coordinates can be accomplished using the ArcMap component of ArcGIS (ESRI, 2015). Coordinate data of site locations are maintained in the geodatabase identified by ReachID and can be added to an ArcMap project and compared with existing features (park boundaries, USGS Digital Ortho Quarter Quadrangles, National Hydrography Dataset hydrography, previous sample sites) to confirm that coordinate data are valid.

Protocol and Data Processing Level Traceability

The Inventory and Monitoring Division Database Standards (Frakes et al. 2015) requires every datum to be unambiguously traceable to: 1) a specific version of a monitoring protocol, a quality assurance plan (QAP), and suite of standard operating procedures (SOPs) and 2) the QA/QC dependent data processing level (e.g., raw, provisional, certified) as defined in Section 5 *Data Processing Levels* section of the Minimum Implementation Standards for Network Projects (Frakes and Kingston 2017).

Data processing levels provide a standardized way of defining the degree that QA/QC has been performed on a dataset. Once a dataset has passed the QA/QC procedures specified in the protocol and standards as identified in the QAP, the dataset will be considered “certified” as defined in Section 5 *Data Processing Levels* section of the Minimum Implementation Standards for Network Projects (Frakes and Kingston 2017, Table 2).

The database contains tables that document the assigned data processing level and identifies the protocol narrative, SOPs, and QAP used for each sampling period. This makes the assigned methods traceable to the record level through the Sampling Periods (Figure 11).

As versions change, the database is updated to reflect a new collection of narrative, quality assurance plan, and SOPs. This provides a chronological record of the specific versions of the protocol narrative, individual SOPs, and Quality Assurance Plan followed for each sampling period throughout monitoring of the Vital Sign.

Table 2. Data processing certification levels based on the documented quality assurance and quality control procedures applied to a data set. (taken from Minimum Implementation Standards for Network Projects, Frakes and Kingston 2017).

Code	Label	Summary	IsActive	Process Order
R	Raw	Data in their state of original collection, either in the form of paper or electronic field forms, electronic sensor files, or digital media (photo, video, or audio). These data generally have been collected following some QA procedures, but have not undergone any QC procedures.	Yes	10
P	Provisional	Data that have undergone all planned QA procedures, have been entered into the system of record, and have passed verification procedures to ensure that data in the system of record match source data (field forms, sensor data files, etc.). Provisional data have not undergone all QC procedures.	Yes	20
A	Accepted	Data that have undergone all QA and QC procedures as defined in a published protocol and associated SOPs. Used in cases where a Quality Assurance Plan has not been developed and approved.	Yes if no QAP	30
PC	Pre-Certified	Data within a system of record that have undergone defined minimum levels of QC as defined in the QAP, rendering the data suitable for use for a limited set of designated or defined purposes only. The QC procedures that have been performed are communicated clearly to the end user. In general, this status should only be used in cases where data are routinely used to support park management decisions before certification can be completed.	Yes if QAP	40
C	Certified	Data have been through all QC procedures as outlined in the published protocol and the associated published QAP. Data either meet the documented quality standards, or data that do not meet the standards are appropriately identified at the observation level; procedures for data collection, processing, and transformation have been documented in appropriate metadata formats; data are suitable for use.	Yes if QAP	50

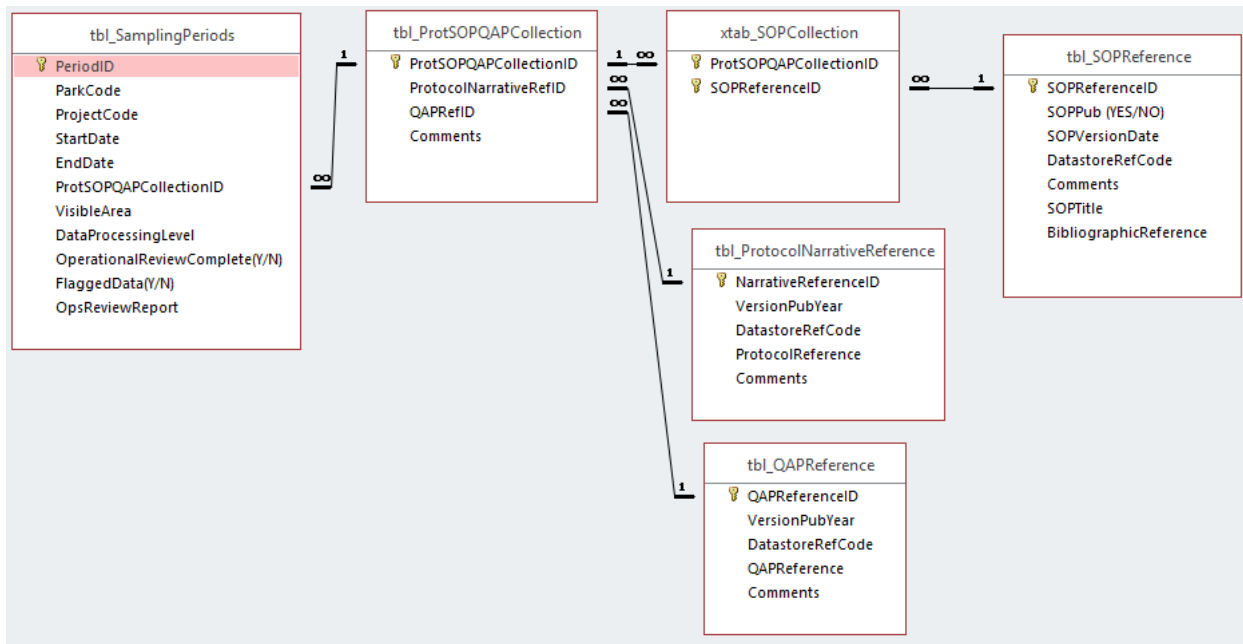


Figure 11. Database tables and their relationship for documenting the specific versions of the protocol narrative, individual SOPs, Quality Assurance Plan, Data Processing Level, and Operational Review Status followed for each sampling period.

Operational Review Status

Additional IMD guidance requires annual operational review (see Reporting SOP) for all active monitoring protocols (Mitchell et al. 2018) to identify the following:

1. Suggested protocol improvements
2. Protocol deviations implemented in the field

3. Approved protocol revisions based on operational review results
4. Protocol associated safety risks identified/verified

Status of the completion of an operational review, whether any data were flagged, and a link to the review report are also traced to the Sampling Period in the database (see Figure 11).

File Organization

The various databases, reports, and GIS coverages used and generated by HTLN create a large number of files and folders to manage. Poor file organization can lead to confusion and data corruption. As a standard data management technique, files pertaining to the project are managed in their own folder with subfolders: Analysis (for data analysis), Data (for copies of archived data as well as data sheets), Documents (for supporting materials related to the project), and Spatial info (for various spatial data and GIS coverages). The databases are managed in designated folders that contain current, archival, and legacy versions in subfolders.

Version Control and Audit Reporting

Secure data archiving, version control, and audit reporting is essential for protecting data files from corruption. All monitoring databases are stored in a designated location on the server. Prior to any changes to a dataset (data entry, minor/major design modifications), steps are taken to allow for tracking of changes over time (versioning control). Proper controls, communication, versioning, audit reporting, and read/write permissions helps ensure that only the most current version is used in any data analysis and that historical “certified” data remains unchanged.

Databases are copied from the server before modification and renamed with an incremental version number. Major version changes include migrations across software versions and complete rebuilds of front-ends and analysis tools. Minor version changes include bug fixes in front-end and analysis tools. Each major version is assigned a sequentially higher whole number in the far left position (e.g., “2.0.0”). Each minor version is assigned a sequentially higher middle number (e.g., “2.1.0”). Versioning due to data entry is assigned a sequentially higher right number (e.g., “2.1.1”).

Once a dataset has passed the QA/QC procedures specified in the protocol and standards as identified in the Quality Assurance Plan (under development), the dataset will be considered “certified”. The Project Manager is responsible for archiving the previous version, placing the “certified” version back on the server, and approving its release to the general public.

Data Publication

Annually, in fulfillment of the Data Analysis and Reporting Requirements (Gallo, K. memorandum dated 4/23/2018), the dataset will be uploaded to IRMA DataStore by the Data Manager. The dataset is flagged as “read only” for all users except the Program Manager, Data Manager, and Project Manager. A metadata database is developed using ESRI ArcCatalog 9.x and the NPS Metadata Tools and Editor extension and will be available at the NPS I&M GIS server (<http://science.nature.nps.gov/nrdata/>). Water quality data will be stored at the EPA

STORET National Data Warehouse and be publicly accessible via the Internet. Additionally, data requests can be directed to:

Heartland I&M Network
Attn. Data Manager
Wilson's Creek National Battlefield
6424 W. Farm Road 182
Republic, MO 65738-9514
417-836-5313 or 417-732-6438

References

Frakes, B., S. Kingston, and M. Beer. 2015. Inventory and Monitoring Division database standards: September 11, 2015. Natural Resource Report NPS/NRSS/NRR—2015/1035. National Park Service, Fort Collins, Colorado.

Frakes B and Kingston S. 2017. Minimum Implementation Standards for Network Projects v.3.0. National Park Service. Last accessed 5/23/2018:
<https://irma.nps.gov/DataStore/DownloadFile/588637>

Gallo, K. 2018. Data Analysis and Reporting Requirements memorandum dated 4/23/2018. Last accessed 5/23/2018: <https://irma.nps.gov/DataStore/DownloadFile/600325>

Mitchell, B., A. Chung-MacCoubrey, J. Comiskey, L. Garrett, M. MacCluskie, B. Moore, T. Philippi, G. Sanders, and J.P. Schmit. 2018. Inventory and Monitoring Division protocol review guidance. Natural Resource Report NPS/NRSS/IMD/NRR—2018/1644. National Park Service, Fort Collins, Colorado.

National Park Service (NPS). 2016. Certification Guidelines for Inventory and Monitoring Data Products. Fort Collins, CO.

National Park Service (NPS). 2016. Integrating QA/QC into the IMD Pipeline. Fort Collins, CO.