

Title: Illuminating the NARS data entry black box: what happens between sample collection and data availability for use in assessments?

Authors: Karen A. Blocksom, David V. Peck, and Teresa K. Magee

Abstract:

The steps between field collection of data and samples and availability of the resulting data from National Aquatic Resource Surveys (NARS) can appear to be a black box. This presentation is intended to shed some light on that process. The pathway for data depends on their source and type. Certain pieces of information about samples come directly from field forms, and the process of scanning these forms and extracting data from them involves a great deal of error checking to ensure the integrity of the data. In addition, biological assemblage data (i.e., fish or plants) recorded directly on field forms undergo extensive reconciliation of names before any statistical analysis can occur. However, data from samples which are sent to laboratories for analysis, including water chemistry and chlorophyll, for example, require additional processing. These sample data are first matched back to the sampling visit ID in the database, then checked for holding time and detection limit issues. For water chemistry data in particular, a series of calculations and plots are then used to identify potential outliers. By the time the data are made available for use in analyses, they have been through a thorough quality assurance (QA) process. To help ensure the transparency of the process, the database is designed to accommodate the tracking of all changes to data made during all QA evaluation steps. For many study components, including but not limited to physical habitat, benthos, fish, and periphyton, a series of metrics is then calculated for use in statistical analysis. Once various condition indicators have been developed and various thresholds determined, they are applied to sample data to assess site condition. This entire process is formalized into R scripts, from the initial error-checking stages to metric calculation to determining condition based on thresholds. Eventually, we hope to compile all of these scripts into an R package for distribution and use by states and other entities engaged in biological assessment.