QUALITY ASSURANCE PROJECT PLAN

FOR THE DEVELOPMENT AND MAINTENANCE OF THE GREAT LAKES REGIONAL DATA MANAGEMENT SYSTEM

FINAL | September 22, 2017

Submitted to: Ben Shorr, Annie Gibbs, Robb Wright - Project Managers (NOAA Office of Response and Restoration)

Developed by: Dawn Smorong (Maven Environmental), Ann Jones and Courtney Arthur (Industrial Economics, Incorporated), and Michael Tweiten (Exa Data & Mapping, Inc.)

This page intentionally left blank.

Table of Contents

Acronyms

AOC	Area of Concern
DIVER	Data Integration, Visualization, Exploration, and Reporting
DMS	Data Management System
DMT	Data Management Team
ERD	Entity-Relationship Diagram
ETL	Extract, transform, load
GLRI	Great Lakes Restoration Initiative
NOAA	National Oceanic and Atmospheric Administration
ORR	Office of Response and Restoration
QAPP	Quality Assurance Project Plan
QA/QC	Quality Assurance/Quality Control
USEPA	U.S. Environmental Protection Agency

1.0 Project Introduction

This Quality Assurance Project Plan (QAPP) provides an overview of the processes the National Oceanic and Atmospheric Administration's (NOAA) Office of Response and Restoration (ORR) uses to assemble natural resource damage assessment (NRDA)-related response, assessment, and restoration data, as well as historical data collected from hazardous sites around the country. This QAPP will provide an overview of the processes used to incorporate data in NOAA's Data Management System (DMS) and outlines the assigned responsibilities for quality control activities associated with each segment of the process.

The primary objective of the QAPP is to ensure integrity during the data acquisition, manipulation, conversion, and translation of environmental data collected in the Great Lakes to a format that may be distributed by NOAA. In order to ensure that the compiled data is accurate, relevant, and comparable (to the original source data set), each segment of the process includes well-defined and documented quality control activities and descriptive documentation.

1.1 Purpose

As part of the Great Lakes Restoration Initiative (GLRI), ORR is continually expanding its existing Great Lakes Data Management System (DMS) to create a centralized repository for environmental data collected throughout the Great Lakes basin. This DMS provides tools to support management actions and restoration planning that will expedite the development, implementation, and monitoring of cleanup and restoration projects throughout the Great Lakes Basin. As federal, state, and local agencies, non-governmental agencies, and private citizens depend on this central data repository as a source of high-quality data to support decision-making, it is essential to standardize procedures for checking, processing, and maintaining the data to ensure data integrity.

1.2 Background Information

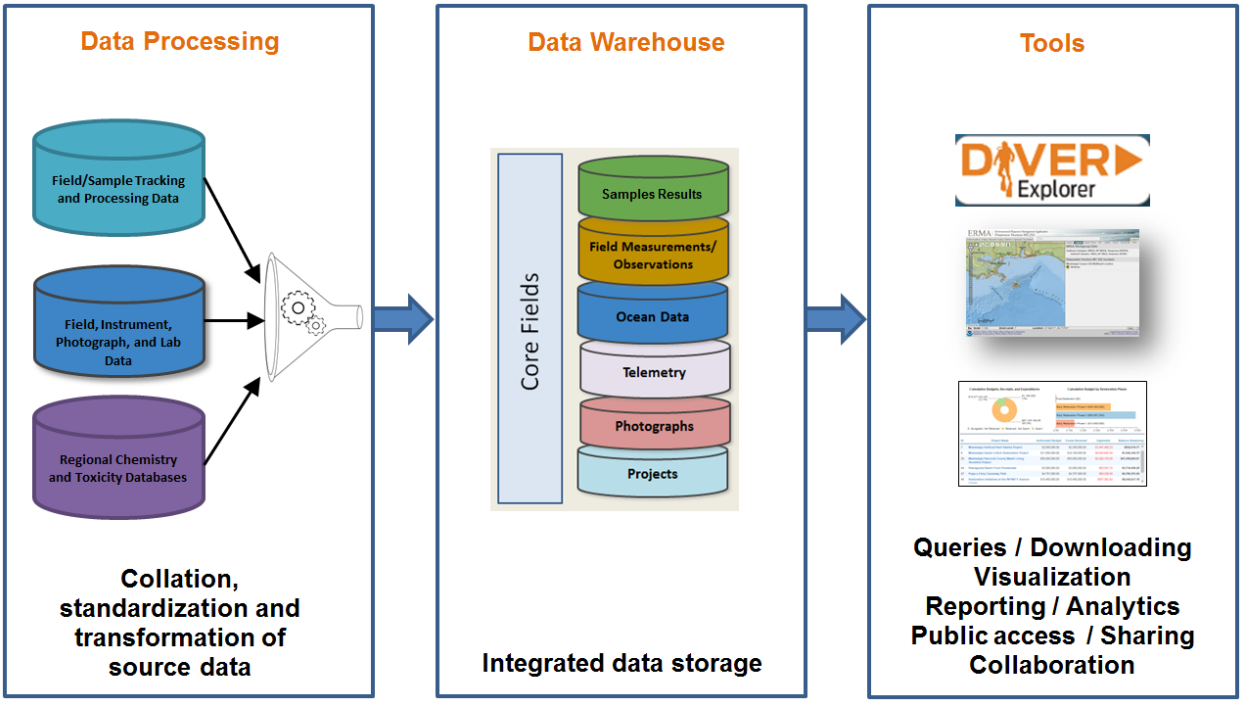
Access to environmental data is essential to the cleanup and restoration of Areas of Concern (AOC) where beneficial use impairments[1] are driven by environmental contamination and degradation. Often, project site data are from multiple sources and are not readily comparable due to differences in measurement units, summation methods, constituent lists, treatment of measurements below analytical detection limits, depths at which sediment and water samples are taken, types of biotic tissues analyzed, or datum and projection of coordinate data. The existing NOAA Data Management System provides a centralized repository for Great Lakes environmental data that facilitates aggregation and comparison of data across studies and sites and helps overcome the "data silos" problem often encountered in business and scientific data, where related data sets of different type and structure exist in isolation from each other.

[1] https://www.epa.gov/great-lakes-aocs/beneficial-use-impairments

NOAA's Data Management System includes three main components (**FIGURE 1**) for managing and analyzing the data:

1. **Data Processing Systems**, for incorporating:
    - Chemistry and toxicity data into NOAA's Great Lakes Regional Chemistry/Toxicity Database,
    - Other data types into DIVER (e.g., telemetry and oceanographic data), and,
    - Field/sample tracking and processing information.[2]
2. **Data Warehouse**, for storing and integrating the different incoming data types;
3. **Tools,** including:
    - DIVER (Data Integration, Visualization, Exploration, and Reporting), NOAA's web-based application for integrating and distributing NRDA-related response, assessment, and restoration data, as well as historical data collected from hazardous sites around the country.
    - ERMA (Emergency Response Management Application), NOAA's web-based mapping system which is integrated with DIVER.

Figure 1. NOAA's Data Management System



The Great Lakes Regional Chemistry/Toxicity Database is a relational database that is based on a standardized database structure, which can be used to manage and access surface and sub-surface sediment, soil, and water environmental-contaminant chemistry data; sediment and water toxicity data; tissue contaminant-residue chemistry data; and oil/tarball chemistry data. While the database is focused

[2] Data processing procedures for field/sample tracking and processing information are not described in this QAPP, as the Great Lakes Data Management System does not regularly utilize this data pathway.

on data from the aquatic environment, data from associated upland areas are also stored in the database. The database system accommodates information from multiple studies and combines field sample collection information with analytical chemistry and toxicity results. The database was designed to minimize data redundancy and organize data in a structure useful for data archiving, data analysis, GIS analysis, and other uses.

The DIVER application serves as a warehouse and portal for data related to the Great Lakes region (among other regions). DIVER is designed to address data volume and integration requirements associated with the GLRI project. In addition to ingesting the regional Chemistry/Toxicity databases ('Sample Results' cylinder in **FIGURE 1**), DIVER brings in field observations and survey results, photographs, telemetry, and data from oceanographic instruments. In addition, DIVER may be used to store original workplans and other related files such as project-specific study plans, data collection methods, quality assurance documents, field definitions and data dictionaries, and related management plans. These files may be uploaded directly to DIVER File Collections alongside data files pertaining to analytical chemistry, field observations, photographs taken during sampling trips and other field research, instrument data, and telemetry data.

DIVER is designed for flexible queries and data processing. The DIVER Explorer tool allows the user to construct custom queries using a series of filters that winnow the data down to a subset of interest. The user can display results on a map, on charts, and in tables, all of which are interactive.

DIVER is a proven, powerful data management and data delivery system that facilitates data evaluation and discussions conducted by a wide range of stakeholders. Users can easily select, explore, and summarize data using pre-programmed queries that sort and analyze the data in a variety of ways, such as comparing site data to Sediment Quality Guidelines, and readily depicting areas of potential concern. The program automatically displays the query results on a map or the data can be readily exported in CSV, XLS, XML, SHP, and KMZ formats, for use with other data analysis and mapping software. Graphical displays through NOAA's Emergency Response Management Application (ERMA) facilitate a regional approach to assessment and remediation by showing the spatial relationships of contaminant, toxicity, resource, and land use data.

The Data Management System, as a whole, has proven to be a useful tool to compile, analyze, and display contaminant data to support a variety of remedial and restoration decisions. The data are routinely used at Superfund and Great Lakes Legacy Act sites in investigation and sample plan design, ecological risk assessment, cleanup level derivation, remedial alternative development and evaluation, mitigation strategy development, sediment and soil remediation design (including time-critical removal actions), natural resource damage assessment (pathway evaluation, injury determination, restoration project development, and scaling damages), and long-term effectiveness monitoring programs and other management activities needed to cleanup and restore the Great Lakes ecosystem.

All agencies involved in management actions and restoration in the Great Lakes region will benefit through increased volume and accessibility of the data as well as the technical input provided by the coordinating agencies. This effort also provides consistency with the U.S. Environmental Protection Agency's (USEPA) National Sediment Inventory database, which uses the base structure of the NOAA Chemistry/Toxicity database. Continued development of the Great Lakes Regional Chemistry/Toxicity Database also provides partner agencies with a powerful outreach tool that can be used to share information with the public in an easy to use format.

All acquired data sets are subjected to a screening and preview process to ensure the respective data sets include the appropriate matrix-specific data elements prior to processing the data into the Data Management System. Data sets are then binned into priority categories. Data sets are next translated to a NOAA Template (Template) format. Data in the Templates undergoes additional Quality Assurance/Quality Control (QA/QC) checks using the NOAA Template Tester (Template Tester) to ensure consistency with the database rules. After the data set has been tested and corrections have been made, the revised data is converted to a standardized format and the data is subject to quality control procedures that validate the accuracy of the data transformation. The data set is then incorporated in DIVER and a final quality control check is performed to ensure the data are correctly represented within the system.

1.3 Organization of the QAPP

This QAPP includes the following sections, which provide an operating framework for:

- Roles and responsibilities for developing and maintaining the Data Management System
- Process and activities required to incorporate data into the System
- Quality control procedures and check points for each of the critical activities
- Overview of data management activities
- Procedures for updating this QAPP document

Content from the Great Lakes Watershed Environmental Database Project QAPP (NOAA 2011) has been adapted to generate this document.

2.0 Roles and Responsibilities

For the purpose of this QAPP, roles are described for external data providers (Track A and Track B) and the NOAA Data Management Team (DMT).

A distinction is made between external data providers and data collected by NOAA, as more comprehensive data management processes are utilized by the NOAA DMT to support a broad spectrum of the NRDA data management processes, including: field planning; sample intake; field sample information management; laboratory analyses; tissue processing; and analytical data validation (DV). These processes are not described in this document.

2.1 External data providers

The Data Management System will be further developed by acquisition of data and information from external data providers such as the USEPA Great Lakes National Program Office (GLNPO) sediment assessment grants program; USEPA Superfund, Resource Conservation and Recovery Act (RCRA), and Water programs; State sediment programs; U.S. Army Corps of Engineers (USACE) sediment programs; U.S. Geological Survey (USGS) National Water Quality Assessment (NAWQA) program; and other programs as identified through the project.

Historically, the NOAA DMT utilized an informal process to work with data providers to process data into the NOAA Chemistry/Toxicity database format. This involved acquiring and reviewing data provided in a variety of formats, and resolving missing or incomplete data prior to processing. Under this model (termed 'Track A' data providers), the bulk of the QA/QC tasks involved with data processing were the

responsibility of NOAA's DMT. NOAA will continue to work under the Track A data provider model using this informal system.

Under a pilot project with the agencies involved in the St. Louis River AOC [Minnesota Pollution Control Agency (MPCA) and the Wisconsin Department of Natural Resources (WDNR)], NOAA has assisted with developing alternate procedures for processing data. Under this model (termed 'Track B'), data providers are delegated specific, and more comprehensive, QA/QC tasks. The St. Louis River AOC stakeholders are utilizing DIVER as their primary Data Management System and the goal of the data processing procedures is to ensure that the quality of the data incorporated into NOAA's DMS meets the group's standards for data consistency and quality. More information about the data flow and QA/QC process for the St. Louis River AOC stakeholder group are detailed in a flow diagram developed by NOAA, MPCA and WDNR.[3]

This QAPP will generally describe the data processing and QA/QC procedures utilized for Track B data providers, as NOAA is interested in following, or adapting, this model to accommodate agencies involved in other Great Lakes AOCs.

2.2 NOAA Data Management Team (DMT)

NOAA's Data Management Team is responsible for implementing the procedures described in this document. The DMT consists of NOAA staff and contractors who maintain the Data Management System for managing and analyzing the data, including NOAA's Chemistry/Toxicity database and DIVER. While numerous researchers and contractors assist NOAA in the data management effort, most of the data management processes described in this document are supported by Industrial Economics, Incorporated (IEc) and Exa Data & Mapping Services (Exa).

All project participants are responsible for following all procedures and QA/QC practices as stated in this QAPP. Specific responsibilities for project activities are outlined in **TABLE 1**.

2.3 Training

NOAA's DMT provides general support on the use of the Data Management System. This support may occur through scheduled webinars or through direct telephone conversations between the DMT representative and technical staff. In certain cases, members of the NOAA DMT may conduct a formal online or in-person training session related to the use of certain elements of the Data Management System. These training sessions are customized for specific participants, and the material covered varies, depending on training needs and participant preferences.

All members of the DMT must complete training on the Templates and Template Tester prior to assisting with data processing tasks. Training on other elements of the Data Management System, such as uploading documents to the File Collection or reviewing posted data, are dependent on specific tasks assigned to DMT members.

[3] The Great Lakes DIVER Data Flow and QA/QC Process Map for the St. Louis River AOC can be accessed at the following link:
https://www.diver.orr.noaa.gov/documents/20233/53415/SLRAOC+Data+Flow_and_QAQC+Process+Map.pdf

Table 1. Roles and Responsibilities for NOAA's Data Management Team

Role	Responsibility
Prepare QAPP	Exa and IEc
Monitor implementation of QAPP	
Coordinate and prepare changes and additions to the QAPP	
Review and approve QAPP	NOAA-ORR:
	Ben Shorr, *Project Manager*
	Robb Wright, *Project Manager*
	Annie Gibbs, *Project Manager*
	Rebecca Held Knoch, *NOAA GLRI Coordinator*
Review data set priorities and progress	Ben Shorr, *Project Manager*
Reset priorities, when necessary	
Develop/maintain data processing tools and associated documentation for data bound for the NOAA Chemistry/Toxicity database	Exa Data
Process and implement QAQC procedures for data bound for the NOAA Chemistry/Toxicity database	
Develop/maintain DIVER data processing tools and associated documentation	Industrial Economics, Incorporated (IEc)
Process and implement QA/QC procedures for data bound for direct ingest to DIVER	
Update information in Tracking systems	All project participants

3.0 Overview of Tasks Involved in Data Processing

3.1 Introduction

Great Lakes environmental data are ingested into DIVER following one of three major pathways, depending on the data type. A different Template is used to process data for each pathway, as follows:

- PATH 1 - ChemTox Template – stores analytical chemistry data (for multiple matrices; sediment, water, tissue, oil, etc), and toxicity data
- PATH 2 - BioLab Template – stores sample-based non-chemistry laboratory data (which may be paired or unpaired with contaminant chemistry data)
- PATH 3 - FieldObs Template – stores field observations, measurements, and surveys (not sample based; may be paired or unpaired with contaminant chemistry data).

See **TABLE 2** for additional information on the Templates.

A major distinction between the data processing pathways is that Path 1 data is stored in the NOAA Chemistry/Toxicity database prior to being served over the DIVER interface, while Paths 2 & 3 data are

ingested directly to DIVER (i.e., this data is not stored in the NOAA Chemistry/Toxicity database structure).

Table 2. Description of the NOAA Templates for Data Bound for DIVER

PATH	Descriptive Name	Name for Reference	Example Data	File Name
1	Chemistry/Toxicity Results	ChemTox	Sediment chemistry Tissue chemistry 10-day Hyalella survival toxicity test	NOAA_Template_ChemTox_Excel_V2.9_20170831.xlsx NOAA_Template_ChemTox_V2.9_20170831.accdb
2	Biological and other non-chem laboratory analysis (sample-based)	BioLab	Benthic invert community metrics Fish histopathology Fish deformities Stable Isotope analysis	NOAA_Template_BioLab_V1.0_20170831.xlsx
3	Field measurements/ biological surveys (not sample-based)	FieldObs	Bird counts Vegetation surveys Percent cover estimates Water quality parameters	NOAA_Template_FieldObs_V1.0_20170831.xlsx

Each of the data pathways have similarities and differences in the data processing steps that ultimately lead to raw data being ingested and served over DIVER. The data processing steps are described in detail in the following sections, and where there are differences in the processing steps depending on pathway, these will be specifically noted. See **FIGURE 2** for an overview of the data processing steps for each path.

The process to convert data from the original format to the DIVER system includes the following steps. Each of these steps or activities requires specific checks and/or documentation to ensure that step is error free and integrity of the original data has been maintained.

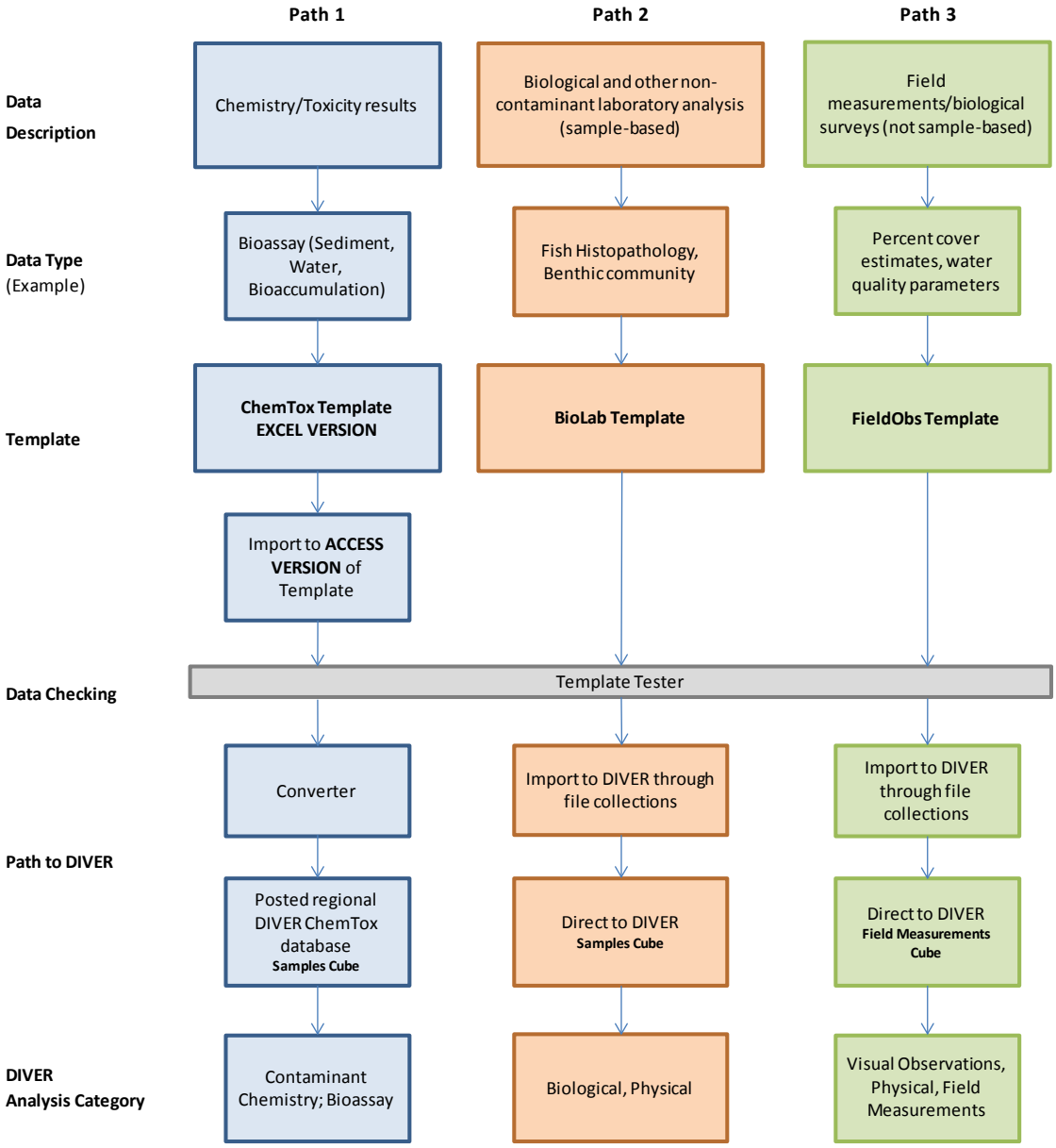
1. Identification and acquisition of candidate data sets;
2. Maintenance of data set information in NOAA Tracking System (Path 1) or DIVER File Collections (Paths 2 & 3).
3. Screening* and initial prioritization of data sets;
4. Preview* of data sets; acquisition of additional documentation, if necessary;
5. Population of Template from source electronic data files;
6. Run QA/QC checks via Template Tester and resolve issues;
7. Conversion of data to NOAA Chemistry/Toxicity database format (**Path 1 only**);
8. Upload to master NOAA Chemistry/Toxicity database, including executing automated QA/QC

routines, then initial (not public) upload to DIVER (**Path 1 only**);

9. BioLab/FieldObs Template – initial (not public) upload to DIVER (Paths 2 & 3 only)
10. Final QA/QC checks in DIVER
11. Final upload to DIVER Portal (password protected) for review
12. Data with a sharing status of "Publicly Available" (i.e. not provisional or unvalidated) published on the Public DIVER website monthly.

**Note that the screen and preview process for data bound for the ChemTox Template is more formalized than that of the BioLab and FieldObs Templates.*

Figure 2. Data Processing Overview



The following paragraphs provide an overview of each of the data processing steps, including the QA/QC procedures followed by the NOAA DMT.

3.2 Description of target data and acquisition

NOAA and its Contractors collaborate with partner agencies to ensure that the NOAA Chemistry/Toxicity database has current versions of data from major data collection programs within the Great Lakes region (e.g., Great Lakes Mussel Watch; GLSed database). In addition, NOAA data managers additionally work with federal, state, and local agencies to identify additional data that are necessary to support AOC management actions and move the AOC towards delisting.

Key stakeholders, such as the St. Louis River AOC group, prepare and maintain their own inventories for candidate data sets to include in DIVER. Data sets received from these groups will be in Template format and will have passed the Template Tester checks, allowing for a streamlined process for incorporating into DIVER.

TABLE 2 provides a description of the types of data that are captured in a standardized format in the three NOAA Templates. Certain data types, such as air quality information, are not supported for standardization through the NOAA Chemistry/Toxicity database due to limited historical use in remedial and restoration activities. As needed, additional data types can be stored in DIVER under Paths 2 and 3 or as related data files. Unstructured information from field studies may also be submitted (e.g., study reports, photographs, GIS shapefiles), which may be stored in the DIVER File Collections (see **Section 5.3** and **Section 5.5** for additional details) and will be accessible to end users.

3.3 Tracking data sets

Path 1

The NOAA Tracking System is designed to track all data sets bound for the NOAA Chemistry/Toxicity database. This system provides the status of individual studies and data sets, in terms of priority, responsibilities for processing, and status of processing. This system is shared online and provides a mechanism for members of the DMT and NOAA managers to quickly access information on current data sets.

Paths 2 & 3

The DIVER File Collections provide a mechanism to track the status of data that will be processed into the BioLab and FieldObs Templates and will be ingested directly to DIVER. As a data set is marked for inclusion into DIVER, a member of the DMT or the data provider will start a new File Collection where the NOAA Template will be uploaded for inclusion in DIVER. See **Section 5.3** and **Section 5.5** for more information on the DIVER File Collections.

3.4 Path 1 - screening, prioritization of candidate data sets, and previewing

Screening

The purpose of the screening process is to conduct a quick assessment of the general information included in the data set and assign primary (“HIGH”), secondary (“MEDIUM”), or tertiary (“LOW”) priority for subsequent data conversion. For data bound for the ChemTox Template, general information about the data will be entered into the Screening Form so that a preliminary priority can be assigned. An example of the Screening form is included in **APPENDIX A**.

Prioritization

A data set will receive a "HIGH" priority if the following criteria are met:

- The data set includes at least one of the main data types (chemistry, toxicity, sample-based non-chemistry data, field observation data) and these data are available in electronic format;
- The target data collections were governed by a quality assurance program and the data were collected by or under the oversight of a partner agency, and project-specific QAPP guidelines were met;
- The data are necessary to make management decisions in the AOC or for NOAA-specific project such as a NRDA case;
- The data were collected mostly within the Great Lakes Basin; and
- The data are relatively recent, with the exception of sediment core data.

If a study was governed by a quality assurance program, but the project specific QAPP guidelines were not met, the data may still be incorporated in DIVER, but this data quality concern will be clearly indicated in the DIVER interface through documentation in the data management system (e.g. record level data quality attributes or study notes).

If there are gaps or unknowns, then the data set will receive a MEDIUM Priority (until the data gaps can be resolved). Data designated as HIGH or MEDIUM priority must be received in electronic format. Data sets that are received only in hard copy format will be automatically designated as LOW priority.

Data will generally be processed in the order received, beginning with all data sets with a HIGH priority designation. Data sets may be further ranked for processing based on some or all or the following criteria, if a processing backlog is encountered. For example, additional considerations include:

- Large, diverse, or otherwise "important" data sets will get higher priority--this would be determined with NOAA's partners;
- Data received from key stakeholder groups, which submit data in Template format, passed by the Template Tester will get higher priority, as they can be processed into DIVER more readily;
- Data sets with multiple sample types collected may be of higher priority (e.g., a data set with chemistry and toxicity data may have higher priority than a data set with only chemistry data);
- Data sets with a variety of chemical classes measured (e.g., samples with multiple chemical classes may be of higher priority than samples with only metals quantified), and/or availability of QA/QC data and/or validated data may be of higher priority;
- Data sets for which timely access is provided to documentation and files will get higher priority (e.g., a high priority data set will be delayed if data acquisition is delayed); and
- For some data sets, the Screener will use professional judgment.

The NOAA Tracking System will capture the priority and the status of individual data sets. Status entries include: In Review; In Process; Ready for BE*; Awaiting Estimate Approval; On Hold; Cancelled; Complete. All studies in the Tracking System are considered "In Queue," however studies with some

status designations (i.e., On Hold) may never be incorporated in DIVER if screening criteria are not satisfied).

**Ready for BE indicates that the data set is ready to append to the NOAA Chemistry/Toxicity database.*

Preview

The purpose of the preview process is to identify data gaps, inconsistencies, and problems with the data set. An example of a data Preview Form is also provided in the materials in **Appendix A**.

3.5 Paths 2 & 3 - screening, prioritization of candidate data sets, and previewing

Data provided as part of Paths 2 & 3 are expected to be screened by the data provider to determine the data quality, whether the data conform to study-specific quality assurance documentation, and the relative priority of data sets. Given the wide variety of sample-based non-chemistry data as well as field observational and measurement data that may be provided in Paths 2 & 3, the DMT has not produced guidelines for prioritization of these data and instead sets an expectation that the data provider will screen and prioritize candidate data according to their standards and needs. That prioritization will be communicated to the DMT when the data set is provided. However, data sets that are paired with high-priority chemistry and/or toxicity data will receive a higher priority, due to the need to standardize the relationships among these data sets on inclusion to DIVER.

Consistent with the previous process described above for Path 1 data, the purpose of the process for Paths 2 & 3 is to identify data gaps, inconsistencies, and problems with the data set once it has been included in DIVER. The DMT will preview the data in DIVER and conduct tests to ensure data quality and integrity. These are discussed in more detail below.

3.6 Populate Templates, Tester QA/QC checks

Population of the Templates

The ChemTox Template is provided in both Microsoft™ Access and Microsoft™ Excel formats, and is designed for users to manipulate and enter original data sets into a database structure that will allow efficient conversion to the NOAA Chemistry/Toxicity database structure. The ChemTox Template can store multiple 'studies', which is defined as a set of samples collected from a site during a period of one year (i.e., generally, a sampling event). Note that a 'data set' may have multiple studies with different priorities. A 'data set' is defined as a set of files provided by one agency or institution.

The BioLab and FieldObs Templates are provided in Microsoft™ Excel format, and are designed for users to manipulate and enter original data sets into a structure that will allow efficient ingest to DIVER. The BioLab and FieldObs Templates are designed to store only one study per template file.

All three Templates contain a series of data tables that are more flexible than that required by the NOAA database structure and DIVER. They are intended to capture all of the data in as close to the original format as possible, and to be flexible for the data providers.

There are no key fields enforced in the Templates; rather, the Template Tester is used to find problems in relationships between tables, duplicate records, and other issues. Refer to the NOAA Data Processing QAPP (in process) for details on the template and conversion process and a description of the Template database structure.

Template Tester Tool

The Template Tester tool is a Microsoft™ Access database and VBA application that has been designed with the objective of identifying errors and omissions in the completed Template files (Paths 1, 2, and 3). The ChemTox template undergoes the most rigorous checking by the Template Tester, as the BioLab and FieldObs Templates are meant to be more flexible and have a more direct path to be presented in DIVER.

The Template Tester is a QA/QC tool that will check for internal consistency and returns detailed error messages to the user that identifies issues that are relatively straightforward for the user to repair.

There are three primary categories of checking routines included in the Template Tester, which are applied to all three Template types (ChemTox, BioLab and FieldObs):

- Check required fields and relations: checks several key elements (e.g., entries in required fields; table relationships are maintained);
- Unique records: test to ensure that each record is unique; and
- Additional Checks: conducts a variety of additional checks to ensure that content is valid and database conventions are adhered to.

In addition, the following checks are also applied to ChemTox Template data:

- Checks to toxicity data that are not covered in the previous checks, and,
- Checks to qualifier codes that are not covered in the previous checks.

When a check is executed, the Template Tester generates queries and reports to display the results. Errors that are identified as "critical" must be resolved before submitting the Template file(s) for further processing.

Refer to the NOAA Data Processing QAPP (in process) for guidance on the use of the Template Tester and the detailed List of Checks conducted by the Template Tester.

3.7 Conversion and upload to NOAA Chemistry/Toxicity Database

Conversion to NOAA Chemisty/Toxicity database format (Path 1 only)

The Converter is also a Microsoft™ Access VBA application, which converts the QA/QC checked and revised ChemTox Template data into the required database field types, formats, and standardized codes. The following data standardizations and calculations are conducted during this step:

- Sums for certain parameters (e.g., PAHs, PCBs) are calculated
- Bioassay statistical significance is calculated
- Units are converted to the database standard
- Depths are converted to the database standard
- Qualifier codes are standardized
- Standardized codes are applied, as specified in the dict schema (e.g., valid values, species, chemical codes, tissue type)

Refer to the NOAA Data Processing QAPP (in process) for details on data standardization and calculation procedures.

This step includes management of the NOAA database dictionary files (e.g., chemdict, testdict), as conversion of data sets often requires the addition of new chemical codes, species codes, or other codes. The Converter is used to transform each set of data into Access tables compatible with the NOAA database. **FIGURE 3** presents the entity-relationship diagram (ERD) for the NOAA Chemisty/Toxicity database.

During the Conversion process, a high level of QA/QC review on each new data set occurs to ensure that the imported data match the original (provided) data and to identify potential problems (e.g., missing values, inconsistent units, or outliers) with the original data. A series of QA/QC checks are also performed to identify problems with internal consistency and errors that may have been introduced during the conversion of data from ChemTox Template format to database format. Refer to the NOAA Data Processing QAPP (in process) for specific checks implemented during the Conversion stage of processing.

Upload to NOAA Chemistry/Toxicity database (Path 1 only)

When a conversion is complete, the study is appended to the Regional Chemistry/Toxicity database. During this appending process, several additional data preparation and data checking steps are completed.

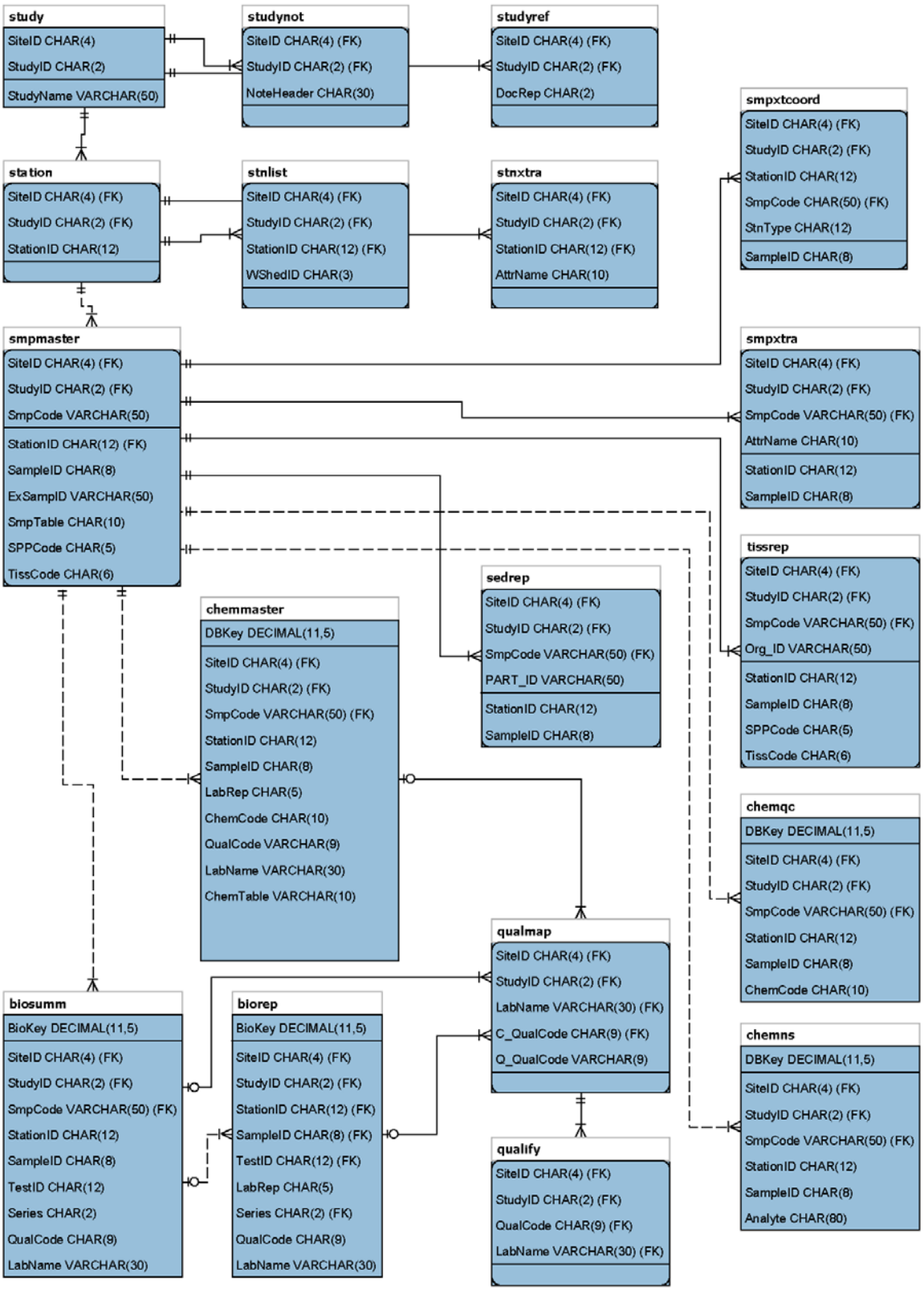
First, the DMT ensures that no duplicate sample or station records are inadvertently appended.

In addition, the DMT members edit and finalize study-related documentation that is stored in a Study Notes table. Within the table, information is captured on sample ID coding, data management decisions, and data handling performed during the conversion. Any calculated sums are defined (e.g., percent fines calculations from percent silt and percent clay values). If the data being appended are from a new study, the data manager creates new records for the Study Notes table. If the data being appended are for an existing study then the existing Study Note record is updated.

When these steps are completed, the DMT member runs the final QA/QC routines to ensure that the data to be appended conform to the database integrity rules including, but not limited to, the following checks:

- Ensuring each record in the database tables is unique based on the database rules (e.g., unique records in the Station table based on SiteID+StudyID+StationID primary key);
- Ensuring all records of “child” database tables have “parent” records in sample, station, and study tables;
- Ensuring all coded fields match accurately to codes and definitions in dictionary tables (e.g., use appropriate species codes for tissue samples);
- Ensuring that units in the chemistry table are standardized based on units defined in the chemistry code/analyte name dictionary;

Figure 3. NOAA Database Entity-Relationship Diagram



- Ensuring that fields critical to correctly appending, processing, and ingesting the data have the correct valid values;
- Ensuring bioassay data tables are correctly coded for sample type and statistical significance codes; and
- Ensuring all calculations done in processing and database preparation are accurate and consistent.

If any of the QA/QC routines identify errors, the DMT member investigates and corrects problems, ensuring adherence to critical database relationships and rules. When all tests are completed the data are appended to the Regional Chemistry/Toxicity database on the DMT member workstation. The master Regional Database schema is then updated and the updated database is ingested into the DIVER data warehouse on a nightly basis, and available through the DIVER Explorer application.

3.8 Initial upload to DIVER (Paths 2 & 3)

For Paths 2 & 3, the data provider should provide data in the standardized NOAA BioLab or FieldObs Template. The data should conform to the standards and guidelines set forth in this document, the NOAA Data Processing QAPP (in process), and any trainings and communications with the DMT. The Template will be uploaded to a DIVER File Collection for the study. The data will then be processed through the Template Tester to identify errors and omissions, and to ensure correct identification of any data sets that have paired data (e.g., chemistry/toxicity data) and should therefore undergo an additional layer of standardization across templates. The DMT will review the file and make further corrections, as needed, following a set checklist to ensure correct transformation and standardization to the DIVER format. When this is completed, the data (in CSV format) will be uploaded to DIVER through the File Collection. Final QA/QC will be conducted as described below.

3.9 Final QA/QC and upload to DIVER

During updates to the DMS, the DMT continually applies routine data confirmation tests and audits to highlight data changes. The DMT reviews the changes to ensure that the data updates have not produced any unanticipated issues. For example, data source providers may attempt to change a field name or make other structural alterations that undermine the DIVER Extract, Transform and Load (ETL) processes. Tests conducted by the DMT may include ensuring the correct number of unique records are visible in DIVER, the spatial representation are correctly displaying, permissions are correctly set such that data are visible to the intended audience, and the correct exports are available in DIVER for end users.

4.0 Quality Objectives and Criteria

Database systems that are developed for project-specific purposes typically establish and enforce overarching data quality objectives that must be met to ensure the successful outcome of the specific project (USEPA 2009). Data quality objectives will specify acceptance criteria for each matrix and measurement and indicate QC samples and/or activities for the indicator. For example, precision, accuracy, and completeness criteria will be set relative to a specific matrix (e.g., water).

For the purposes of the NOAA Great Lakes Regional Chemistry/Toxicity Database, data quality is not assessed specifically against data quality objectives, as the DMS stores data from multiple sample

collection efforts, each of which may have their own specific data quality objectives and acceptance criteria. Thus, the DMT members managing the NOAA DMS are not tasked with validating either field collection information or analytical chemistry results. Where the DMT members identify data gaps or incomplete information, they work with data providers to identify and assemble the most complete and accurate information available. Higher priority is placed on capturing studies that were governed by a quality assurance program and/or were collected by or under the oversight of a partner agency (see **Section 3.4** and **Section 3.5**). Key stakeholder groups (i.e., the St. Louis River AOC group) that collaborate directly with NOAA to capture data of specific quality are responsible for ensuring that field and laboratory data meet QAPP requirements.[4]

Laboratory QA/QC data are not routinely captured in the NOAA Great Lakes Regional Chemistry/Toxicity Database, and although data is not reviewed by the NOAA DMT to assess whether project specific QAPP guidelines were met, information is included in the system to allow end users to assess data quality. For example, an end user may glean data quality from the following information:

- Laboratory and validation qualifiers
- Bioassay qualifiers
- Sampling methods
- Validation level
- Analytical methods
- Analytical Detection Limits/Reporting Limits (DL/RL)
- Quality Control (QC) Batch and Sample Delivery Group (SDG) information
- References to multiple project documents (sampling plans, validation reports, etc.)
- Study notes/metadata
- Files including QA/QC results can be stored in the File Collection area of DIVER (**Section 5.3** and **Section 5.5**) and are accessible to end users.

5.0 Data Management

5.1 NOAA Chemistry/Toxicity database structure

The NOAA database structure combines field sample collection information with analytical chemistry and/or toxicological results reported by the laboratories. The DMT seeks to assemble all the information as accurately as possible, using the information provided by data providers and the results reported by the analytical and toxicological laboratories once the chemical data had been appropriately validated.

The objectives for the NOAA Chemistry/Toxicity database included combining the information gathered from the field and results of the chemical/toxicity analyses into a relational database management system designed to minimize data redundancy and effectively organize data in a structure useful for a variety of needs such as data archiving, data analyses, and use with GIS. **FIGURE 3** illustrates the structure of the database. The database structure has a five-tier hierarchy, i.e., five major table types that are split

[4] The Great Lakes DIVER Data Flow and QA/QC Process Map for the St. Louis River AOC can be accessed at the following link:
https://www.diver.orr.noaa.gov/documents/20233/53415/SLRAOC+Data+Flow_and_QAQC+Process+Map.pdf

into a relational structure. The five types include the study table, station table, sample tables, chemistry tables, and bioassay tables. **TABLE 3** describes the content of each table type.

NOAA Chemistry/Toxicity Database Rules and Specifications

Data captured in the NOAA database adhere to the following rules and specifications:

- For consistency and compatibility with legacy systems (based on an Xbase format), the NOAA Chemistry/Toxicity database tables are created with a structure requiring that the key fields used to link related tables have matching field sizes and the content of these fields must match between tables, in terms of upper and lower case lettering. The decisions regarding field sizes were made in consultation with NOAA staff; refer to the NOAA Data Processing QAPP (in process) for specifics on field sizes in each of the database tables.
- In the station table, each record of the table is unique, based on SiteID + StudyID + StationID. Furthermore, within a study, each unique set of coordinates (as latitude/longitude) must have a unique StationID. No two StationIDs for the same study may have the same coordinates. Two sets of coordinates expressed as decimal degrees (the database standard format) are deemed to be the same when they match after rounding to six decimal places.
- If two or more samples of the same matrix (e.g., sediment) are collected for the same study at the same location (lat/long coordinates) on the same day at the same depth and subjected to similar chemical analyses, the reviewer will assign them the same StudyID + StationID and root SampleID; however, one of the two samples will have a "D" suffix assigned to the SampleID to designate that sample as a field duplicate (samptype = "FDUP").
- If two tarball samples are collected from the same vegetation sample at the same time, it is not possible to know if they attached onto the vegetation at the same point in time. Therefore, the tarball samples collected from the same vegetation will get different SampleIDs.
- If two or more organisms of the same species are collected for the same study at the same location (lat/long coordinates) on the same day, and they are assigned different field SampleIDs, and they were not combined as a composite sample, they will be assigned two different sample IDs.
- If two or more samples share the same matrix, study, location (lat/long coordinates), day, and depth, but are subjected to different chemical analyses, the reviewer will assign them the same StudyID + StationID and SampleID. Thus, two field SampleIDs may be merged into a single Sample ID so that all chemical analyses are associated with a single sample record in the sample table.
- As noted, a lab may split one field sample into different components, according to species, size, etc. In the laboratory EDD, the different components are usually distinguished by a suffix added to the original client sample ID. Within the NOAA Chemistry/Toxicity database, the two resulting samples may be assigned different sample IDs.
- A suffix may be added to a SampleID to relate a sample that has been split into different fractions or components. Some samples include the following:
    - Vegetation samples may be rinsed at the lab and the rinsate and vegetation components analyzed separately. Analytical results are provided in weight/weight units

(e.g., µg/kg). Both samples are assigned the same StudyID and StationID. The rinsed vegetation sample (e.g., sea grass) is assigned a SampleID of T001 with the tissue type identified as Leaves. The rinsate is assigned a SampleID of T001R with the tissue type identified as External Material.

- Crab samples may be split into component parts. All samples are assigned the same StudyID and StationID. The component parts are assigned SampleIDs of T001H (Hepatopancreas), T001L (Gill), and T001M (Muscle).
- Water samples may be filtered and analyzed for the dissolved portion (passing through filter) and particulate portion (trapped on filter and resuspended in deionized water). The filtered portion is assigned a suffix of "F" and the particulate portion a suffix of "P," yielding the SampleIDs W001F and W001P, respectively.

5.2 Common data model in DIVER

To organize data in a consolidated framework, DIVER data managers defined a series of overlapping data models that integrate the breadth of environmental characterization data collected through response, assessment, and restoration phases of natural resource damage assessments, as well as data from historical sites around the country. These "common data models" cover physical samples gathered for further laboratory analysis, field observations and recordings, photographs, telemetry information from cetaceans and turtles, oceanographic data, and reference information on additional videos and analysis. DIVER data managers continually review and adapt the models to integrate additional information. More information about the common data models and their specifications are detailed in the NOAA DIVER Environmental Data Specification document (NOAA 2017).[5]

An established "common data model" can be shared with data providers (e.g., field researchers) and data managers. It provides a collective blueprint for aligning definitions and meaning, and defines templates for exchanging information. The common data model concept is flexible and scalable. As new or different information is received, data managers can expand the schema to accommodate the new information.

A key objective of DIVER is to accommodate the querying of sample data along with associated non-sample data (e.g., field measurements, continuous-read instruments, photos). To pursue this objective, DIVER data managers identified the overlapping concepts generally implicit in each data set and defined the valid value list for promoting consistency between each data category. The overlapping fields in each of the common models represent our core data and required elements for all data sets. **TABLE 4** lists the core fields within the common models. The core information makes the related data available for searching and download. If a specific core data field is not applicable to a particular data set, it is assigned a default value (typically "Not Defined") so that comprehensive data searches return full results.

[5] The NOAA DIVER Environmental Data Specification can be accessed at the following link: https://www.diver.orr.noaa.gov/web/guest/data-overview

Table 3. Database Table Types and Descriptions (bold text indicates main data tables; other tables are supplementary tables)

Table Type	Description
study	**The study table provides basic information regarding the study (e.g. name, contact, etc.) and identify the multiple sample collection events. Each study is assigned a unique, two-character StudyID, which is used to link to tables in the other tiers of the database hierarchy.**
studynot	Contains information regarding the document(s) associated with the study and data.
studyref	Contains study-specific meta-data for specific topics.
station	**The station table identifies locations for samples that were submitted for chemical and/or toxicological analyses. Each record of the table has a unique combination of SiteID + StudyID + StationID. Stations are defined for each study by a unique set of geographic coordinates reported as latitude and longitude.**
Stnlist	Contains a list of stations in each Station Group inlcuding historical Query Manager Watersheds
stnxtra	Contains additional attribute data for stations.
smpmaster	**The sample tables provide information about the samples collected for chemical and/or toxicological analyses, including collection date, depth (if relevant for the matrix type), and sample type (e.g., field sample, field duplicate, composite sample). The master sample table stores all matrix types. Each record within the sample tables is unique based on SiteID + StudyID + StationID + SmpCode.**
smpxtcoord	Contains additional coordinates associated with a sample, for example composited sub-sample locations.
smpxtra	Contains additional attribute data for samples.
tissrep	Sample information for part samples that make up composited tissue samples.
sedrep	Sample information for part samples that make up composited sediment samples.
chemmaster	**The chemistry tables store the results for chemical analyses, for all matrix types. Supplementary chemistry tables store additional information related to analytical chemistry results. Each record is unique, based on SiteID + StudyID + StationID + SampleID + Labrep + Chemcode. Chemcodes are ten-character codes assigned to analytes. Using chemcodes eliminates the potential confusion associated with the multiple ways in which an analyte name might be written (e.g., dibenzo(a,h)anthracene versus dibenzo[a,h]anthracene) or with chemical synonyms used by different laboratories (e.g., 2-methylphenol versus o-cresol). Different Labrep codes are used for results where a duplicate chemical record might otherwise occur in the chemistry table. For example, if a sample was analyzed by the same analytical method and two different laboratories, the results may be distinguished by Labrep.**
chemqc	Stores quality control samples, such as field blanks, that are not included in the chemmaster table.
chemns	Stores Tentatively Identified chemicals (TICS) and originally reported sums that are not included in the chemmaster table.
biosumm	**Mean of sediment bioassay results, with one record per sample tested.**
biorep	Contains replicate data from the sediment bioassay results.

Table 4. DIVER Environmental Data Specifications – Core Fields

Field Name	Field Definition	Field Set Within DIVER Explorer	Field Value Source
Case/Activity	The name of the case incident or the activity used to collect data.	Case/Activity Overview	User-Generated
Collection Workplan	The workplan under which the field data were collected.	Case/Activity Overview	User-Generated
Region	Region	Case/Activity Overview	User-Generated
Workgroup	The Technical Working Group under which the field data were collected.	Case/Activity Overview	User-Generated
Workplan Topic Area	The main resources of focus of a Collection Workplan.	Case/Activity Overview	User-Generated
Workspace Name	Name of the Portal Workspace where data were entered.	Case/Activity Overview	User-Generated
Collection Form	The type of the data submission form used by the field team to submit raw field data.	Collection Summary	User-Generated
Collection Study Name	The name of the study under which the field data were collected.	Collection Summary	User-Generated
Data Category	General category of data collection (e.g., Instruments, Photographs, Samples, or Visual Observations).	Collection Summary	User-Generated
Data Classification	The purpose for which data was collected within the case incident or activity.	Collection Summary	User-Generated
Data Source	The originating owner of the dataset.	Collection Summary	User-Generated
Source Type	General owner/source of the data (e.g., NRDA, Response, Responsible Party).	Collection Summary	User-Generated
Collection Matrix	The type of sample or record collected (e.g., Sediment, Water, Photograph, Wipe).	Field Data	User-Generated
Sample ID	Unique ID assigned to each sample by the field sampler.	Field Data	User-Generated
Station/Site	Station or site identifier. This is often defined by the workplan and/or recorded by the field team, but may be standardized to database requirements.	Field Data	User-Generated
Date	Data collection date, as year, month, and day.	Location/Date/Time	User-Generated
End Latitude	End Latitude	Location/Date/Time	User-Generated
End Longitude	End Longitude	Location/Date/Time	User-Generated
Start Latitude	Start Latitude	Location/Date/Time	User-Generated

Field Name	Field Definition	Field Set Within DIVER Explorer	Field Value Source
Start Longitude	Start Longitude	Location/Date/Time	User-Generated
State	The state where the field event took place.	Location/Date/Time	User-Generated
Analysis Category	General category of analysis performed (e.g., Plankton_Nekton, Visual Observation, Contaminant Chemistry). For additional detail, see Analysis Type and/or Analysis.	Results: All Data Types	User-Generated
Analysis Status	Status of samples in the analysis process as reported by laboratories or through results (e.g., Archived, Results Available, In Analysis Queue etc.).	Results: All Data Types	User-Generated
Analysis Type	Subcategory (i.e., type) of analysis performed, such as Biomass, Hematology, Genetics, etc. For additional detail, see Analysis.	Results: All Data Types	User-Generated
Review Status	Extent of data quality review performed.	Results: All Data Types	User-Generated
Sharing Status	Identifies extent of data distribution (e.g., Publicly Available).	Results: All Data Types	User-Generated
Region ID	Region ID	Case/Activity Overview	DIVER-Created
Station Group List	Predefined sets of grouped stations/locations	Case/Activity Overview	DIVER-Created
DIVER Dataset	DIVER's internal database table name	Collection Summary	DIVER-Created
File Collection ID	Record identifier for the corresponding DIVER file collection.	Collection Summary	DIVER-Created
Record ID	Identifier for each observation data sheet entered into the DIVER database.	Collection Summary	DIVER-Created
Trip ID	Identifier for tracking field collection events and the way data files were provided to the Data Management Team (one Trip ID per file collection or zip file).	Collection Summary	DIVER-Created
Image Id	Record identifier for a particular photograph.	Results: All Data Types	DIVER-Created
Link to Related Files	Link to source files for related data	Results: All Data Types	DIVER-Created
Photo URL - Midsize	Mid-sized image	Results: Photographs	DIVER-Created
Photo URL - Original	Original image	Results: Photographs	DIVER-Created
Photo URL - Thumbnail	Thumbnail sized image	Results: Photographs	DIVER-Created
QM Site ID	Identifier for a site in the Query Manager database.	Results: Samples	DIVER-Created

5.3 DIVER related data

Based on a needs assessment conducted by NOAA in 2012, the majority of Great Lakes data is expected to fit into either the Samples or the Field Observations data category. Other types of data may be considered for inclusion into DIVER on a case-by-case basis. Data providers have the ability to upload supporting files to DIVER through the File Collections or by submission to the DMT, and associate those files with a particular project. Supporting files may include structured or unstructured data, thereby allowing users to upload and access different types of data which may not match an existing DIVER data category. These data would then be available for download in a native format, but would not be available for querying in DIVER Explorer.

5.4 Documentation and Records Management

5.4.1 Data set documentation

Documentation for each data set will be compiled by the DMT member responsible for each respective data processing step. All files will be submitted by the DMT member to the final repository and will be maintained for as long as the NOAA Chemistry/Toxicity database is maintained. The initial repository for these files will be with ORR and possibly with NOAA's National Center for Environmental Information (NCEI), which is NOAA's official archive. References and/or links to the following types of data set documentation, if available, will be compiled and added to the data set metadata within the NOAA database, and stored within the File Collections area of DIVER:

- All quality assurance documentation for the original data set including QAPP, validation reports, etc.;
- Laboratory analytical reports; and
- Final project reports summarizing the data.

In addition, completed QA/QC documentation for all steps of database population, as described in this QAPP, including Screening Forms, Preview Forms, results of Template Tester, and results of Data Conversion checks will be maintained.

5.4.2 Technical documentation

Technical documentation for database development and data translation routines will be stored and maintained in the Database Program files.

5.5 Method for serving data (DIVER)

The DIVER Data Warehouse can accept both structured and unstructured data. NOAA typically uses File Collections within the DIVER Portal application to gather unstructured field collected information including scanned field forms, field notes, photos, GPS coordinates, and other related files. DIVER File Collections can also accept structured (spreadsheet) data in its original format, as well as data that have been transcribed to match a data model that maps the data source fields to corresponding DIVER fields. Electronic data deliverables (EDDs) have been developed for contaminant chemistry, bioassay, and field-collected observations and are available for download by data providers. Data providers should provide metadata in ISO 19115-2 format, or sufficient information to create these metadata records, with the

expectation that metadata files will be available for download with any original data files. For more information see the DIVER Data Specification (NOAA 2017).[6]

Operationally, the DIVER Data Warehouse works with data providers to create ETL procedures from their data warehouse or data store. DIVER data managers work with data providers to review the data and determine the best course of action to integrate the data into the DIVER data models. This process can vary in complexity depending on the format of the incoming data and the frequency of updates.

An ETL process is the standard approach for creating a unified data warehouse suitable for advanced querying and analysis. To combine the various sources into a cohesive data set, the DMT uses Pentaho data integration tools as a core part of the DIVER system.[7] The DIVER ETL process includes a series of import routines for extracting and translating source data into the DIVER common data model structures, standardizing incoming data feeds, and loading into the DIVER data warehouse. **FIGURE 4** illustrates the process for data integration.

The first step in the DIVER update process is to collect, or feed, new source data records into the system in their original format. This process is scheduled to run nightly, but can also be initiated at any time by a DIVER Administrator. Source information comes in a variety of file-based formats, including MS Excel, MS Access, and CSV files. An authorized individual can upload data sources that are file-based to the DIVER File Collections, which can be imported to the DIVER data warehouse using data templates. Alternatively, data can be pulled live from external databases maintained in management systems such as PostgreSQL, MySQL, Oracle, or MongoDB. DIVER can reach out to such external systems over protocols such as HTTPS or FTPS. Regardless of the source, the initial goal is to create local copies of the source tables in DIVER's core ETL processing database. This database runs on PostgreSQL, an established database management system designed for secure storage and easy data retrieval by multiple applications. During the update process, each of the previously integrated source tables is removed and reloaded with the current version. These tables provide the raw data that feed the rest of the process.

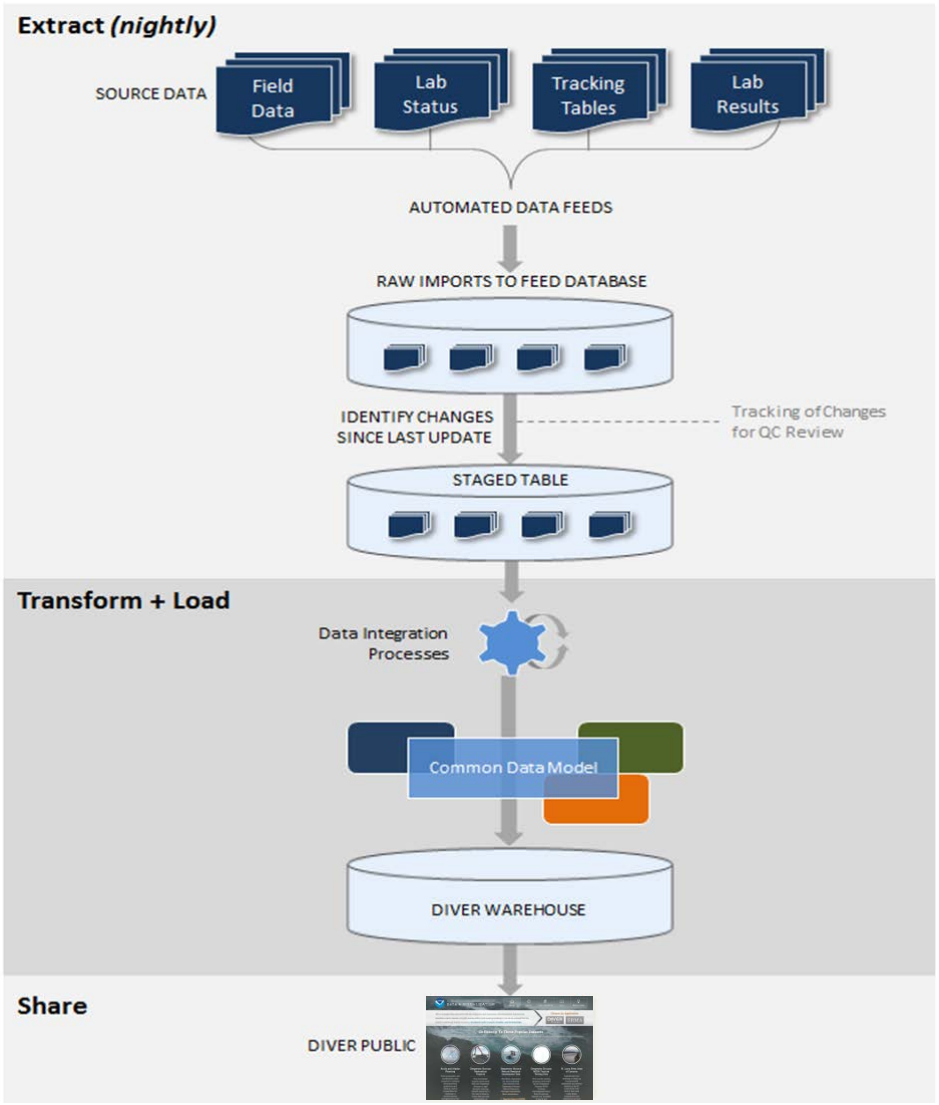
The next step in the DIVER update process examines the source data sets for changes since the previous load, and then stages the information for creating the final DIVER data sets. The DMT employs a set of mirrored tables for each of the source tables, allowing comparison of the previous staged state and the updated feed state. The stage tables are then updated incrementally to capture any changes made to existing records, remove records no longer present, or add new data records.

Through this approach, the DMT can track the last update date for each record. The DMT also maintains a table-specific tracking of inserts, updates, and deletes to support quality-control reviews. Based upon the nightly summary, the data team can quickly monitor whether new data were entered or whether a source system update was missing.

[6] The DIVER Environmental Data Specification may be accessed at: https://www.diver.orr.noaa.gov/web/guest/data-overview

[7] http://www.pentaho.com/; Data Integration, Big Data, Analytics

Figure 4. Process for Data Integration



During this stage, DIVER also conducts simple data cleanup. For example, any leading or trailing spaces are deleted from each field value to ensure the data are properly displayed and links can be made between disparate sources. DIVER then executes a series of ETL tools that merge the disparate data into the unified, sample-specific data set. The tools also apply a series of data standardizations to promote data consistency, including the following:

- Apply consistent capitalization (e.g., change matrix from water to Water);
- Apply spelling corrections for species and other dimensions;
- Standardize different versions of the same value (e.g., FL, Fla, Florida); and
- Assign new attributes (e.g., if Lab = ABC then Analysis Category = XYZ).

At the conclusion of the DIVER build process, the system pushes the analytic data set to a column-oriented database designed specifically for high performance querying.

DIVER EXPLORER

The DIVER application includes a custom query tool called DIVER Explorer that provides a customized data access tool. DIVER Explorer contains tools that allow the user to interact with complex data across different data types. The query tools provide both pre-set Guided Queries, as well as user-specific custom query capabilities, providing access to hundreds of fields from multiple data categories. Additional tools give the user the ability to drill deeper into the queried data through table filtering, interactive maps, and clickable charts and data displays. Finally, the export tools offer options for exporting queried information as GIS files, spreadsheet data tables, as well as the original data and related files. These capabilities are described below.

Queries

DIVER Explorer provides multiple ways to query and filter environmental data. The query tools pull data directly from the underlying data categories, offering customized summaries, legends, charts, tables, and metadata to provide clear and concise information about the queried data. Two main types of queries include guided queries with a number of pre-selected filters, and broader environmental data queries where the user chooses fields and values to filter.

DIVER Explorer offers dozens of Guided Queries for environmental data. Guided queries allow users to select the type of query, load pre-selected filters, apply additional filters (if needed), and then run the query to display the data. Guided queries are robust, tailored for specific output structures, and contain features including:

- Specialized legends to display contaminant values;
- Multiple contaminant pivot tables;
- Calculations for detection limits and rejected results; and
- Comparisons to sediment quality guidelines.

Environmental data queries are created by iteratively selecting fields and values and successively filtering prior to running the query. DIVER Explorer allows high-level queries (e.g., all data by DWH workplan) but also enables queries at a finer level of granularity (e.g., lab result or observation). Users logged into the DIVER Portal environment can save custom queries for future use, share queries with others, or send queries as an Internet link embedded in an email or a report. DIVER Explorer allows for cross-category querying, eliminating the data silos that environmental data often reside within and helps users identify important relationships in the dataset. For example, queries can be structured to reveal how a given sample matrix is affected at different times or locations. Likewise, queries can be structured to see results for shared attributes such as trips, sites, or stations.

Guided and environmental data queries can be refined using exploratory tools within DIVER Explorer. Initial information about the query is clearly presented in the query filters list at the top of the query page (i.e., the fields and values used to build the query) and on the resulting summary tab when the query has been run. The user can modify what is shown on the web map by toggling the legend. A “Data and Export” tab displays the queried data in a table format, with columns that can be shown or hidden. To further refine the query after an initial run, the user has several options:

- Filter the table tab by values in the fields;
- Draw a polygon on the map to refine data for a specific location;

- View additional information associated with a given row in the table;
- Click on a point in the map to show information in a new tab; and/or
- Click “Edit Query” to edit the query parameters.

Exports

Once a query is finalized, the data can be exported from DIVER for use in other environments. Spatial information can be exported from the map as either an ArcGIS shapefile or a Google Earth KMZ file. The table tab may be customized by the user and exported directly to an Excel table format. All export packages are delivered in a ZIP file with ISO 19115-2 metadata generated based on the specific query details, and may include additional information including study notes and original data packages if relevant.

5.6 Limitations on the use of data

Limitation on the use of individual data sets will be documented using several methods:

- Narrative included in study notes;
- Qualifier code flags on individual results;
- Information on validation level of individual results;
- Within the study report documents included in the DIVER File Collections;
- Within the files containing QA/QC data include in the DIVER File Collections.

If the source data was updated subsequent to acquisition by NOAA, these modifications may not be incorporated in the DIVER system. Users can assess the date the data was added to DIVER by viewing the Upload Date field in the DIVER File Collection for data from Paths 2 and 3 or by reviewing the study notes and database edits report from the Chemistry/Toxicity Database for Path 1 data.

6.0 Review and Update of QAPP

To ensure that the QA/QC procedures described herein remain effective and robust in maintaining data integrity, this QAPP will be reviewed and updated on a regular basis. As the project evolves, NOAA will develop and evaluate systematic remedies and corrections to address any procedural or data errors encountered. Necessary changes to existing procedures as well as the establishment of new procedures that improve the QA/QC process (i.e., improves the accuracy, relevance, and comparability of data) will be incorporated into the QAPP. The ORR Project Manager will be responsible for overseeing the QAPP update and revision process.

7.0 References

NOAA. In Process. Quality Assurance Project Plan for the Development and Maintenance of the NOAA Data Management System. National Oceanic and Atmospheric Administration, Office of Response and Restoration.

NOAA. 2011. Quality Assurance Project Plan, Great Lakes Watershed Environmental Database Project. National Oceanic and Atmospheric Administration, Office of Response and Restoration. QM_QAPP_2011.08.04.pdf

NOAA. 2017. DIVER Application Environmental Data Specification. Version 1.2. September 18, 2017. Available online at https://www.diver.orr.noaa.gov/web/guest/data-overview.

USEPA. 2009. EPA New England Quality Assurance Project Plan Guidance for Environmental Projects Using only Existing (Secondary) Data. USEPA New England, Quality Assurance Unit, Office of Environmental Measurement and Evaluation. Revision 2. EPANESecondaryDataGuidance.pdf

Appendix A – Examples of Screening and Preview Forms (Path 1)

Screening Form

This is the form template to be used when doing the initial screening review for new data sets considered for inclusion in the NOAA Chemistry/Toxicity databases. For questions about how to fill out this form, contact:

Michael Tweiten
Phone: 260-317-6381
michael@exadata.net

Dawn Smorong
Phone: 250-591-1815
dawn@mavenconsulting.ca

Inventory Linkage Information

DataSetID: _____

Date Screen completed: _____

Staff completing Screen: _____

Date of last update: _____

Data Format

Are analytical data available electronically in readily-convertible format? ☐

Data Format Comments: _____

Possible duplicate study or study already in QM? _____

Are there data in the provided reports that are not included in the data files. If so,

describe: _____

Sample Type Information

Are the following sample types included in this study?

- Sediment chemistry (surface) ☐
- Sediment chemistry (sub-surface) ☐
- Soil Chemistry (surface) ☐
- Soil Chemistry (sub-surface) ☐
- Tissue Field Chemistry ☐
- Tissue Lab Chemistry ☐
- Water Chemistry ☐
- Tar/oil Chemistry ☐
- Laboratory toxicity tests (sediment) ☐
- Laboratory toxicity tests (water) ☐

Other (specify): _____

Station Information

Are station coordinates available in electronic format for all stations? ☐

Coordinate system, if known, and other station-related comments:

_____

Sample Information

Number of samples collected (approx): _____

Provide sampling year(s): _____

Provide sampling location (general): _____

Analytical Data

Are results reported with units? _____

Additional Comments _____

Preview Form

This is the form template to be used when conducting a Preview of a data set for potential incorporation of the dataset into the NOAA Chemistry/Toxicity Database. Please fill out as much of the requested information as possible. For questions about how to fill out this form, contact:

Michael Tweiten
Phone: 260-317-6381
michael@exadata.net

Dawn Smorong
Phone: 250-591-1815
dawn@mavenconsulting.ca

***Critical information is designated with an * symbol.**

Inventory Linkage Information

*DataSetID: _____

*Date Preview Completed: _____

*Staff conducting Preview: _____

Date of last update: _____

File and Study Reference Information

*In what electronic format were the data submitted? _____

*Applicable data file name(s): _____

*Is there documentation or a report with the data? _____

*Is the documentation/report available electronically?

*Has the study been published? _____

Provide reference citation: _____

Provide source information for documentation (e.g., URL): _____

Applicable document file name(s):_____

Briefly describe the study objective (e.g., dredged material testing; routine water sampling):

_____

Station Information

*Estimate the % of sampling locations with coordinates: _____

Enter the coordinate system, projection, and datum: _____

Additional comments about the availability and reliability of coordinate information, and if missing coordinates can be determined from a map:

_____

Sampling Information

* Provide sampling dates (mm-yyyy): _____

* Provide sampling location (general) _____

* Are the samples located within the geographic area of interest? _____

Is there a naming convention for samples (especially compositing, field replication) in the Sampling Plan or other documentation? _____

_____

Sediment Sample Information

*Are sediment depths/units of the samples available? Describe. _____

Is information provided for the individual samples in a composite if present (e.g., location, sample time, etc.)? _____

Is there an indication that standard methods were used for the collection, storage and handling of samples, (please specify protocol; e.g., EPA-823-B-01-002 or ASTM E1391, and provide a page reference)._____

Water Sample Information

*Are water depths/units of the samples available? Describe. _____

Is information provided for the individual samples in a composite if present (e.g. location, samples, time, etc.)? _____

Is there an indication that standard methods were used for the collection, storage and handling of samples (please specify protocol; e.g., EPA-823-B-01-002 or ASTM E1391, and provide a page reference).

_____

Were samples filtered or otherwise post-processed in the field or laboratory? Provide details.

_____

Tissue Sample Information

* Is species information provided? _____

Are common and/or scientific species names provided (please specify)? _____

*Is tissue type information provided? _____

Is there information on tissue sample preparation (e.g., scaled, filleted, weighed, homogenized, etc.)?
_____

Is information provided on the length, weight, and sex of the organisms (provide comments)?
_____

Is information provided regarding the preparation of composite samples (e.g., # in composite, tissue type, ranges of weight/length and/or sex/age of individuals included in composite sample)?
_____

Overall Comments for Data Set

_____

_____

_____

_____

Questions on Analytical Methods/Data - enter separate records for each Matrix Type

Questions on Analytical Methods

*Sample Matrix _____

Are preparation/extraction and analytical methods reported (provide comments)?

_____

*Were there analytical laboratory replicates? Are they clearly identified in the data?

_____

*Are all sediment chemistry data reported in dry weight? If not, are %solids or %moisture data provided? Provide comments

_____

*Are all tissue data reported in wet weight? If not, are %solids data provided? Provide comments.

_____

Was total organic carbon and grain size measured on each sediment/soil sample? Data available electronically?

_____

Was the lipid content measured on each tissue sample? Data available electronically?

_____

Are there totals provided (e.g., total PAHs, PCBs, etc.)? If so, is there information on how these values were derived?

_____

Questions on Analytical Data

*Are all results reported with units? If not, then describe the gaps and if available in related reports.

_____

What chemical classes have been measured?

_____

*Are detection and/or reporting limits^ included with the data, especially for data reported as below detection? Describe. _____

*If detection and/or reporting limits^ are not included with the data, are they provided in the documentation (please specify)? _____

Were field QAQC samples collected/analyzed (e.g., field duplicates, field blanks, etc.)? Please specify.

_____

*Does the data set include results for laboratory QAQC samples (e.g., matrix spikes, surrogate analyses, etc)? Is this data available electronically?

_____

Have the data been validated? If so, include details indicating the level of validation (e.g., unvalidated, complete independent review, validation qualifiers included, completeness check, etc.)

_____

Have the data been qualified (e.g., in addition to identifying non-detected results)? If so, are qualifier descriptions available?

_____

Additional Information or Comments:

_____

_____

_____

_____

^Reporting Limit (RL) is the minimum value below which data are documented as nonquantifiable. It is the reporting limit for the sample analyzed, as determined by the laboratory. The Method Detection Limit (MDL) is the minimum concentration of an analyte that undergoes the entire measurement process and can be reported with a stated level of confidence that the analyte concentration is greater than zero. It is the detection limit associated with the method used to analyze the analyte, or parameter, in the sample.

Questions on Biological Toxicity Tests - enter separate records for each Test

Describe the bioassay test information (enter separate records for each Test Description):

*Test description: Duration/Species/Endpoint (e.g.,10-day, H. Azteca, Survival and Growth)

_____

_____

*Test Media (e.g., Bulk Sediment) _____

Test Method(s) (e.g., ASTM 1706) _____

Is replicate data available? _____

*Has biological significance been recorded (e.g., toxic vs not toxic)? _____

Have statistical methods for determining significance been described? Provide details.

Were negative control results provided/documented^? _____

Were reference samples collected and results provided^? _____

Toxicity Data Comments:

_____

_____

_____

_____

^Negative control = lab created "clean" media; reference = field collected sample from probable un-impacted site.