Environmental Data Management in Support of Sharing Data and Management

SCIENCE

Office of Research and Development

FINAL REPORT
ENVIRONMENTAL DATA MANAGEMENT IN SUPPORT OF SHARING DATA AND MANAGEMENT

by

Dr. Lilit Yeghiazarian, Mr. Amr Safwat, Mr. Allen Teklitz, and Mr. Donald Morehead,

Pegasus Technical Services, 46 E. Hollister Street, Cincinnati, OH 45219
School of Energy, Environmental, Biological and Medical Engineering Program, University of Cincinnati, Cincinnati, OH 45221

Dr. Tim Whiteaker and Dr. David R. Maidment

Pegasus Technical Services, 46 E. Hollister Street, Cincinnati, OH 45219
Center for Research in Water Resources, University of Texas at Austin, Austin, TX 78712

Dr. Elly P.H. Best,
Water Supply and Water Resources Division, National Risk Management Research Laboratory,
EPA/ORD, Cincinnati, OH 45268

DISCLAIMER

The U.S. Environmental Protection Agency (EPA), through its Office of Research and Development, funded and managed, or partially funded and collaborated in, the research described herein under Task Order (TO) 0041 of Contract No. EP-C-11-006 to Pegasus Technical Services. This document has been reviewed in accordance with U.S. Environmental Protection Agency (EPA) policy and approved for publication. The views expressed in this report are those of the author[s] and do not necessarily reflect the views or policies of EPA. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. The quality of secondary data referenced in this document was not independently evaluated by EPA and Pegasus.

ABSTRACT

A data management system (DMS) was developed, tested and demonstrated to store and manage water quality and quantity (WQ2) data pertaining to U.S. Environmental Protection Agency/Office of Research and Development (EPA/ORD) research projects in standardized formats. This approach was taken to facilitate accessibility, sharing, integration, and use of this information for simple calculations and inclusion into models by EPA and other users to inform water management decisions. The objectives of this project were: (1) Build a previously identified hydrologic information system (HIS) Observational Data Model (ODM) and convert existing water quality and quantity data, generated by EPA/ORD-research projects, into this format; (2) Develop, test and demonstrate a scalable DMS based on HIS protocols for storage and management of geo-referenced data; and (3) Enable targeted users to access, exchange and integrate geospatially referenced WQ2 information, and to conduct relatively simple calculations and/or run models to inform water management decisions at various watershed scales. The HIS developed by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) was used in this project. The CUAHSI HIS ODM model is widely accepted in the United States, is compatible with the National Water Quality Portal, and is gaining international acceptance as the standard for water data. This HIS is web-based, and uses the WaterML format for sharing hydrologic time series data.

Within the U.S., many organizations measure WQ2 and various biological parameters. Despite the fact that this information is routinely made available to the public, the difficulties in identifying data sources, syntactic and semantic heterogeneity across data formats and metadata make the discovery, access and interpretation of data challenging for research and other stakeholder communities. EPA-ORD seeks to alleviate these challenges and plans to use CUAHSI's HIS and other tools to integrate time series data generated within its local, regional and national watershed projects with data collected by collaborators, and to share these data with other interested parties.

This report presents the results of exploring, implementing and amending the CUAHSI approach to storing and managing data of (1) an ongoing WQ2 monitoring effort in the East Fork of the Little Miami River (EFLMR) watershed in Ohio; and (2) a completed WQ2 monitoring effort in the Shepherd Creek watershed near Cincinnati, Ohio. Data included in the new DMS pertained to physicochemical parameters of discrete water samples, water-immersed sensors, and discrete stream-substrate samples. Data on biological parameters of water-immersed sensors and discrete stream-substrate samples were not included. The functionality of the DMS was demonstrated by estimating an example nitrogen (N) TMDL using a Load Duration Curve (LDC) approach based on data retrieved from the EFLMR DMS and other databases. All data were accessed, explored and visualized through the HIS Hydrodesktop tool.

ACKNOWLEDGMENTS

EXECUTIVE SUMMARY

Many organizations within the U.S. measure water quantity and quality (WQ2) variables such as precipitation, streamflow, water quality parameters (pH, nutrients, etc), and various biological parameters. Results of data analyses are typically published in reports and journal articles. These publications often include technical details of data collection and metadata. Data values are often made available as files that can be retrieved from public websites. Discovery, access and interpretation by research and stakeholder communities are challenging because of syntactic and semantic heterogeneity across data formats and metadata. Pilot projects that render information on WQ2 variables, values, and metadata readily identifiable, accessible, sharable and usable in analytical and modeling applications that accept standardized formats for direct interpretation, can play a vital role in facilitating and accelerating collaborative and environmental modeling activities within EPA-ORD, EPA, as well as between EPA and other agencies and organizations. The pilot project described in this report was conducted to evaluate a web-based, highly standardized and widely accepted hydrologic information system (HIS) for storage and management of EPA-ORD regional watershed WQ2 data.

The main goal of the project was to explore a previously identified HIS for storage of WQ2 data of an ongoing EPA-ORD watershed project, to develop, test and demonstrate the subsequent use of these data for calculation of a total maximum daily load (TMDL), and to demonstrate potential access and use by targeted users. In addition, a wider implementation potential within EPA-ORD was explored.

The HIS developed by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) was used in this project. This HIS provides a database schema called the Observations Data Model (ODM) for consistent storage and management of observations data and associated metadata, including data elements required by the WaterML format. ODM is widely accepted in the United States, it is compatible with the National Water Quality Portal, and is gaining international acceptance as the golden standard for water data. The CUAHSI HIS is web-based and uses the WaterML format for sharing hydrologic time series data.

WQ2 data from an ongoing watershed management research project in the East Fork Watershed of the Little Miami River in Ohio (EFWLMR), pertaining to 2010, served as the initial data source. The DMS work was initiated by downloading and installing a blank ODM database (subsequently called 'DMS') from the CUAHSI-HIS website onto an EPA-ORD server located within the EPA-ORD firewall. Of all data categories explored, the physico-chemical data from discrete water samples, the sensor data from water-immersed sensors for physico-chemical variables, and the physico-chemical data from discrete stream substrate (sediment) samples were committed to the ODM. The biological monitoring data, largely on benthic macrofauna, could not be loaded because no data loader is currently available to which the data may be mapped. The data from the discrete water samples were reformatted and mapped to the ODM Data Loader. Additional columns were inserted into their original Excel sheet for Sources and Sites, and the information required to populate these columns was generated: for Sources, organization/agency collecting the data; for Sites, site-name, latitude, longitude, and elevation. In addition, the column headers were edited to match the ODM vocabulary, particularly the Variable Code. For each existing Excel sheet, a second sheet was added to contain the metadata. After reformatting, all data were committed to the DMS using the ODM Data Loader. The data from the water-immersed sensors were committed directly to the DMS using the ODM

Streaming Data Loader. The data from the discrete sediment samples were also reformatted, and then mapped to a modified version of the ODM Streaming Data Loader. Besides editing and addition of the same information as for the discrete water samples, additional columns were inserted into their original Excel sheet for the sediment categories. The consistency of the data committed to the DMS was validated by a three-step process, including verification of the data category, of the ODM log file, and of the actual data committed to the ODM database (DMS). The same procedure was used to commit water quality monitoring data of the completed Shepherd Creek project to the ODM database, and, thus, also add this new database to the DMS.

The functionality of the DMS was demonstrated with an example nitrogen (N) TMDL calculation using the Load Duration Curve (LDC) approach. Data were retrieved from the DMS and other databases and accessed, explored and visualized through the HIS Hydrodesktop tool. In this case Hydrodesktop accessed the DMS-data via a link to the test publication of the EFW Database for which CUAHSI provides a codeplex sandbox. When implementation of the recently developed DMS is approved for use within EPA-ORD and funding for the DMS to enter into its' 'Production Phase' (involving dedicated server testing, operating and maintenance) is generated, publication of the EFW Database in the CUAHSI HIS Services Central Web Service Registry (and possibly other websites) can be accomplished. The latter step would enable access and use by users in- and outside EPA/ORD, and may greatly facilitate and accelerate collaboration.

CONTENTS

FIGURES

TABLES

APPENDICES

ACRONYMS AND ABBREVIATIONS

ADC	Application Deployment Checklist
cfs	cubic feet per second
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science Inc.
DIN	Dissolved Inorganic Nitrogen
DMS	Data Management System
DMZ	Demilitarized Zone
DWTP	Drinking Water Treatment Plant
EFLMR	East Fork of the Little Miami River
EFW	East Fork Watershed
EPA	United States Environmental Protection Agency (=USEPA)
ERD	Ecosystem Research Division
FDC	Flow Duration Curve
GIS	Geographic Information System
HIS	Hydrologic Information System
LTER	Long Term Ecological Research
Mac	Macintosh
MGD	Mega Gallons per Day
NERL	National Exposure Research Laboratory
NRMRL	National Risk Management Research Laboratory
NWIS	National Water Information System
ODM	Observational Data Model
ORD	Office of Research and Development
OSIM	Office of Science and Information Management
PC	Personal Computer
QA	Quality Assurance
RTP	Research Triangle Park
sa	system administrator
SC	Shepherd's Creek
STORET	STORage and RETrieval system for water quality data
TMDL	Total Maximum Daily Load
TO	Task Order
UC	University of Cincinnati
USACE	United States Army Corps of Engineers
USGS	United States Geological Survey
WQ2	water quality and quantity
WQMB	Water Quality Management Branch
WSWRD	Water Supply and Water Resources Division
WWTP	Waste Water Treatment Plant

1.0:   INTRODUCTION

1.1   Project Background

Within the U.S., many organizations measure water quantity and quality (WQ2) variables such as precipitation, streamflow, water quality parameters (pH, nutrients, *etc*), and various biological parameters. Results of data analyses are typically published in reports and journal articles. These publications often include technical details of data collection and metadata. Data values are often made available as files that can be retrieved from public websites. However, syntactic and semantic heterogeneity across data formats and metadata make the discovery, access and interpretation challenging for research and stakeholder communities.

To address these issues, the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) has developed a web-based Hydrologic Information System (HIS) that uses the WaterML format for sharing hydrologic time series data (www.cuahsi.org). The CUAHSI HIS also provides a database schema called the Observations Data Model (ODM) for consistent storage and management of observations data and associated metadata, including data elements required by WaterML. The CUAHSI model has been widely accepted in the United States and is gaining international acceptance as the standard for water data, with over 150 organizations such as universities, utilities, agencies, businesses, state and local governments. An overview of web-based interactions facilitating the sharing of hydrological data enabled by working with CUAHSI-HIS is provided in Appendix A.

EPA-ORD plans to use CUAHSI's HIS and other tools to integrate time series data generated within its local, regional and national watershed projects with data collected by EPA's collaborators, and to share these data with other interested parties. The *significance* of this effort is that the standardization and integration of data are expected to greatly facilitate collaborative and environmental modeling activities within EPA-ORD, EPA, as well as between EPA and other agencies and organizations in the new Safe and Sustainable Water Research Program's focus areas including Sustainable Water Resource Flows and Sustainable Natural and Engineered Water Infrastructure Systems.

1.2   Project Objectives

This project has three objectives:

Objective 1. Build a CUAHSI ODM and convert existing WQ2 data into the CUAHSI format;

Objective 2. Develop, test and demonstrate a scalable Data Management System (DMS) based on CUAHSI protocols for storage and management of geo-referenced discrete data generated by ongoing watershed management research in the East Fork Watershed of the Little Miami River in Ohio (EFWLMR), as well as newly initiated research at other locations;

Objective 3. Enable targeted users to access, exchange and integrate geospatially referenced WQ2 information, and conduct relatively simple calculations and/or run models to inform water management decisions at various watershed scales.

The results of this project are subject to the Quality Assurance Project Plan ID no W-15812 (Approval date: 11/21/2011).

1.3   Report Outline

The remainder of the report is organized as follows. Section 2 describes data sources used in the project and those projected for use in the future. Section 3 outlines the approach taken by the team to convert the existing WQ2 data into the CUAHSI format. This section meets the first objective of the project. Section 4 describes data publication phases, with a focus on the current, demonstration phase. Section 5 describes the architecture of the DMS developed for the EFWLMR data, and its use for demonstration purposes. Sections 4 and 5 meet the second objective. Section 6 explains how users can retrieve data from the DMS and upload their own, and describes the computer codes developed by the team to inform watershed management decision-making. Section 6 meets the third objective of the project. Sections 7 and 8 identify critical gaps in current approaches and data management and discuss the feasibility of further improvement.

The contents of these sections are briefly outlined below:

2.0   Data Sources
   This section outlines the data sources used in the project and those projected for use in the future.

3.0   DMS Architecture and Use for Exchange and Integration of WQ2 Information
   The CUAHSI-HIS data publication tools are reviewed. The existing WQ2 data and metadata were translated into terminology used by the CUAHSI ODM, these translation methods are outlined in this section. The data were then loaded into ODM. All steps to install the required software and load the data are outlined.

4.0   Data Publication
   Steps needed to publish an ODM database are outlined. Three data publication phases are presented: development, demonstration and production. The data publication is currently in the demonstration phase, whereby a copy of the dataset is housed on a designated server at the University of Cincinnati (UC). This web-service can be accessed by EPA and non-EPA users who have access to the Internet, by searching in CUAHSI HydroDesktop for "EPA - East Fork Watershed" in Ohio.

5.0   Using The DMS: Exchange and Integration of WQ2 Information
   This section describes tools that enable data visualization and analysis. It includes a discussion of HydroDesktop, HydroR and HydroExcel.

6.0   Data Analysis to Inform Decision-Making
   This section contains the description of data retrieval by users from the DMS, uploading data of their own, and the computer codes developed by the team to inform the decision-making process of watershed management.

7.0   Critical Gaps of Current Approaches
This section contains the overview and synthesis of information provided in Sections 3-6, and identification of the critical gaps in the approaches.

8.0   Recommendations for Improving upon Existing Approaches
This section focuses on feasibility assessment for improvement upon the best current approaches and discusses an implementation plan.

2.0: DATA SOURCES

The first source of data is the EPA-ORD/NRMRL's ongoing WQ2 monitoring effort in the EFW. This monitoring program has been in place since 2005. The resulting data include (1) physico-chemical data from water samples; (2) sensor data from water-immersed sensors for physico-chemicial variables; (3) physico-chemical data from stream-substrate (sediment) samples; (4) sensor data from water-immersed sensors for biological variables; and (5) biological data from stream substrate samples. In particular, data collected in the year 2010 were selected for upload into the ODM. All data originating from this data source are subject to the Quality Assurance Project Plan ID No 634-Q-2-0. Approved data may become available for public use approximately 2 years after collection. For example, a 2-year lag-period between data collection and public availability is customary for networks such as the Long Term Ecological Research (LTER) Network.

Other potential data sources are related to projects within the NRMRL Cincinnati Cross-Laboratory Green Infrastructure Study, where water quality monitoring is conducted. Among the latter projects, those on Shepherd's Creek, Quebec Heights, and Pervious Pavement, all located in Cincinnati, OH, served as potential additional data sources.

The monitoring data pertaining to the Shepherd's Creek project, located near Cincinnati, Ohio, served as the second data source. This monitoring program has been in place from 2004 to 2011. The resulting data included physico-chemical data from water grab samples. All data originating from this data source are subject to the Quality Assurance Project Plan ID No S-10386-JA-2-0 (Approval date: 09/02/2011).

**3.0:   DMS ARCHITECTURE AND USE FOR EXCHANGE AND
INTEGRATION OF WQ2 INFORMATION**

In this section the hardware and software requirements to create a HydroServer for data publishing are outlined. The DMS architecture is presented as well as how it can be used for data sharing. This section is organized as follows. The first part, *General Guidelines*, is a condensed guide to steps that need to be taken to publish data through a HIS HydroServer. Detailed step-by-step instructions are given in Appendix A, which is a tutorial from a workshop taught at the University of Cincinnati by Dr. Timothy Whiteaker. The workshop was conducted using a simplified dataset 'RawData', available electronically as part of this report to facilitate user learning. The second part, *Working with East Fork Watershed (EFW) Data*, is a detailed, step-by-step description of work performed with EFW data collected in 2010 and published through CUAHSI HIS. The third part, *Working with Shepherd's Creek (SC) Data*, is a description of the work performed with SC data collected in the period 2004-2011 and published through CUAHSI HIS.

3.1 General Guidelines

3.1.1. Hardware and Software Requirements.

The minimal hardware requirement is a Personal Computer (PC) with 4GB RAM, 500 GB hard disk, 2.7 GHz processor. Macintosh (Mac) users can use the CUAHSI software through a virtual machine.
The software includes:
- Windows 7, XP or Vista
- Microsoft Internet Information Services (IIS) 7 — comes with Windows 7, but may need to be enabled. Enable ASP.NET.
- .NET Framework 2.0 SP2, 3.5 SP1 (free)
- Microsoft SQL Server 2008 R2 (commercial) or SQL Server 2008 R2 Express (free)
  - o Be sure to install the version with SQL Management Studio (aka Database and Management Tools)
  - o Install with these options:
    - – Install with Mixed Mode Authentication.
    - – Specify the 'sa' (system administrator) password that you will remember.
- HIS software (free)
  - o ODM Data Loader - http://his.cuahsi.org/odmData Loader.html
  - o ODM Streaming Data Loader - http://his.cuahsi.org/odmsdl.html
  - o ODM Tools - http://his.cuahsi.org/odmtools.html
- Additional software (not required for HydroServer)
  - o Microsoft Office 2010 (32-bit version)
  - o Google Earth (free) - http://earth.google.com/download-earth.html
  - o HydroObjects (free) - http://his.cuahsi.org/hydroobjects.html
  - o HydroExcel (free) - http://his.cuahsi.org/hydroexcel.html
  - o HydroDesktop (free) - http://his.cuahsi.org/hydrodesktop.html

3.1.2. Review of CUAHSI HIS Data Publication Tools and Service.

The CUAHSI HIS provides Web services, tools, standards and procedures that enhance access to more and better data for hydrologic analysis. HIS software is free and available on the HIS website at http://his.cuahsi.org. A variety of HIS software applications have been built to serve several types of users and scenarios, from data users to data publishers to educators and developers. These include the HydroServer, the ODM, WaterOneFlow, HIS Central, the HydroDesktop. HydroServer is what allows the users to publish their own data. It includes software for publishing observations data with a WaterOneFlow Web service, but it also includes a website and supporting components for geospatial and temporal data visualization. ODM is a data model for the storage and retrieval of hydrologic observations in a relational database. WaterOneFlow is a Web service that facilitates automated and programmatic access to data. HIS Central is a website maintained by the CUAHSI HIS team where users can register their WaterOneFlow Web service. The service then becomes discoverable along with dozens of other Web services already registered with the system, including the services EPA Storage and Retrieval System for water quality data (STORET) and the USGS National Water Information System (NWIS). This makes HIS Central the largest single catalog of the nation's water data. HydroDesktop is a free and open source Geographic Information System (GIS) that allows the user to explore and download Internet data published on HydroServers.

3.1.3. ODM Structure.

ODM is a data model for the storage and retrieval of hydrologic observations in a relational database (Figure 3-1). The purpose for such a database is to store hydrologic observations data in a system designed to optimize data retrieval for integrated analysis of information collected by multiple investigators. It is intended to provide a standard format to aid in the effective sharing of information between investigators and to allow analysis of information from disparate sources both within a single study area or hydrologic observatory and across hydrologic observatories and regions. The observations data model is designed to store hydrologic observations and sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used. A relational database format is used to provide querying capability to allow data retrieval supporting diverse analyses. A generic template for the observations database is referred to as the Observations Data Model (ODM). The specifics of the ODM are documented in Tarboton et al., 2008. The current ODM design specifications can be found in Appendix B of this report.

3.1.4. Installing a Blank ODM Database.

The first step is to download and install a blank ODM database from http://his.cuahsi.org. The ODM database then needs to be attached to the SQL Server. The controlled vocabularies within the database are then updated; this step helps to ensure that the user's terminology is consistent with the peer-reviewed vocabulary of terms maintained by CUAHSI.

3.1.5. Requirements for Input Files to ODM Data Loader: Transforming Data into CUAHSI Format.

User data files need to be transformed into files that are formatted for loading into an ODM database using the ODM Data Loader. The minimum pieces of information, stored in separate files and needed to describe data in the ODM, are:

- Sites
- Sources
- Variables
- DataValues
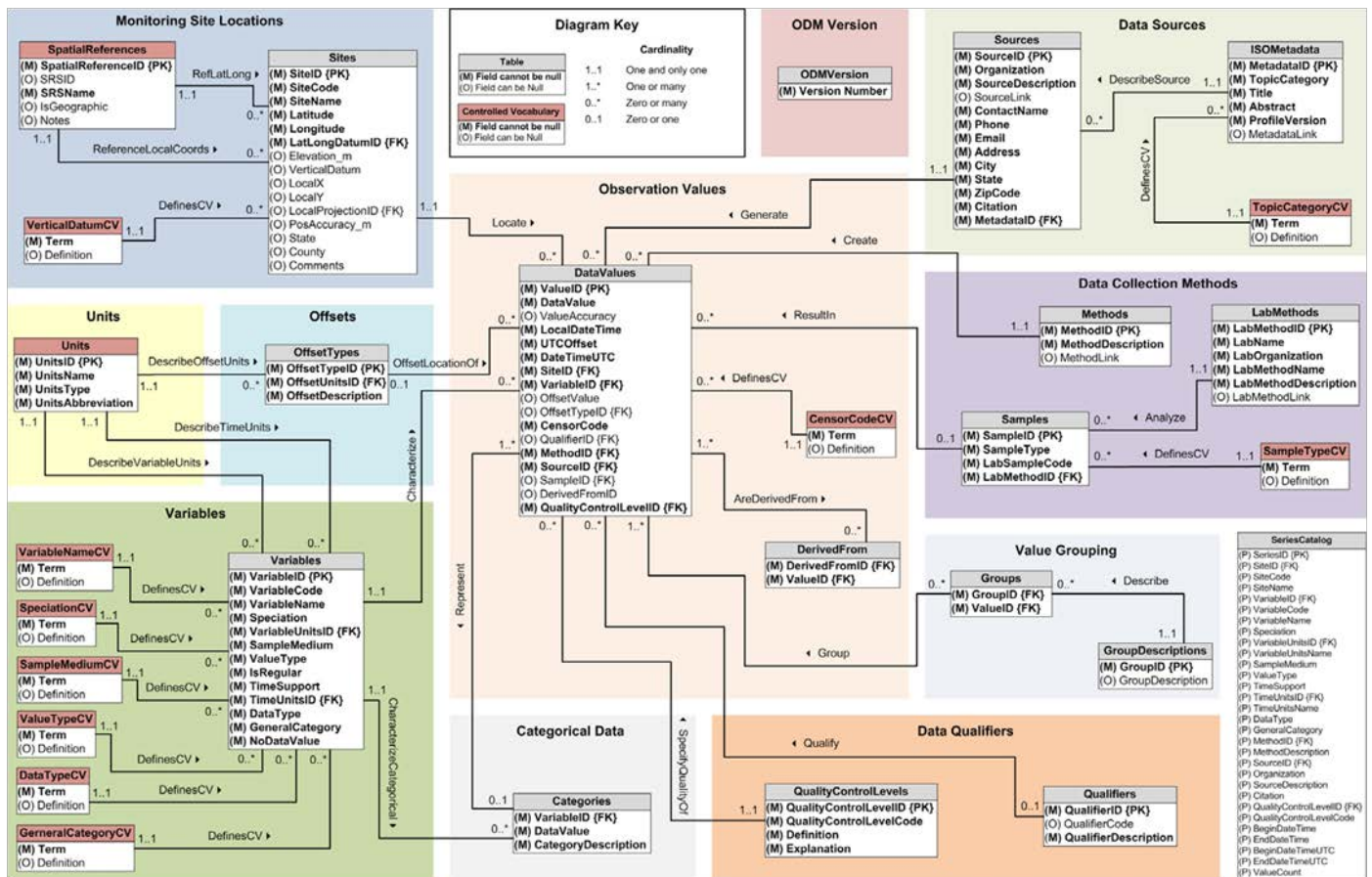
Figure 3-1.   Observations Data Model (ODM) Schema, from Horsburgh et al. 2009

3.1.6. Loading Metadata and Data Values into ODM.

After the *Sites, Sources* and *Variables* metadata and data values are transformed into files acceptable by the ODM Loader, they are *committed* to the ODM database. In addition to data recorded in Excel spreadsheets, text files, *etc*, it is also possible to load data from telemetry systems. Appendix A describes how to use the *Streaming Data Loader* for that purpose.

3.1.7. Working With ODM Data.

Once data are loaded into the ODM database, they can be queried, visualized and analyzed. ODM Tools are used for this purpose. They also allow to derive new time series from data, for example by applying a function to the original data.

3.2. Working with East Fork Watershed (EFW) Data

A list of the data files pertaining to the EFW which were accommodated into the DMS is shown below. The data for the water-immersed sensors for biological variables were not available. The biological data from the stream-substrate samples could not be accommodated because HIS is not programmed to store biological data and it does not carry a biological data loader yet.

1. Physico-chemical data water and sediment
    a. 2010_ESF-EFWS_NutrientData_ODM.xls
    b. 2010_ESF-EFWS_NutrientData_ODM_QC.xls

2. Sensor data from water-immersed sensors for physico-chemical variables
    a. CEC 2009-2010.csv
    b. CLC 2009-2010.csv
    c. FMR 2009-2010.csv
    d. HST 2009-2010.csv
    e. HWR 2009-2010.csv
    f. KRT 2009-2010.csv
    g. LRC 2009-2010.csv
    h. OWT 2009-2010.csv
    i. SHA 2009-2010.csv
    j. SLT 2009-2010.csv
    k. UHL 2009-2010.csv
    l. USR 2009-2010.csv

3. Sediment data
    a. FieldSedimentFractions_8-05_7-09_update91409.xls

All files were formatted to conform with the CUAHSI standard before they were loaded to the ODM database. Most of the column names in those files were renamed.

3.2.1. Description of Physico-chemical Data Water File.

The nutrient data file contained a DATA and a METADATA spreadsheet (Figure 3-2; Figure 3-3). The DATA spreadsheet contained information about sampling date and time, site location, sampling method, variables, and concentrations. Each column in the DATA spreadsheet is explained in detail under the METADATA spreadsheet. The original METADATA spreadsheet has been modified to such an extent that it can be used as source for lookup tables. The new spreadsheet named DataValues has been added to this file. This spreadsheet pulls out all the information from the two spreadsheets DATA and METADATA and creates a CUAHSI

formatted DataValue file. This DataValues spreadsheet was copied to a new Excel file and saved as a csv file.

id $char3.	subid $char7.	pos $char4.	rep 1.	matrix $char2.	type $char3.	analy $char6.	cup 3.	ddate yymmdd10.	dtime time11.	dilf 2.	method $char5.	parea 8.	pht 8.	conc 10.	unit $char6.	cconc 10.
site id (ID)	long id (SUBID)	CRoss id (POS)	REP#	MATRIX	TYPE	Analyte Name (ANALY)	Autosampler-Cup Number (CUP)	Analyte Detection Date (DDATE)	Analyte Detection Time (DTIME)	Manual Dilution Factor	METHOD	Peak Area (PAREA)	Peak Height (PHT)	Peak Concentration (CONC)	UNIT	Peak Concentration (corrected for dilution factor) (ug/l)
QAC	STD	H	1	DI	CAL	TNH4	91	20100106	02:14:20 PM	1	N-1.1	-0.155	-0.00734	0	ug N/L	0.00
QAC	STD	G	1	DI	CAL	TNH4	92	20100106	02:15:25 PM	1	N-1.1	0.25	0.0124	20	ug N/L	20.00
QAC	STD	F	1	DI	CAL	TNH4	93	20100106	02:16:31 PM	1	N-1.1	0.923	0.0472	60	ug N/L	60.00
QAC	STD	E	1	DI	CAL	TNH4	94	20100106	02:17:36 PM	1	N-1.1	1.38	0.0722	80	ug N/L	80.00

Figure 3-2.   Physico-chemical Data Water 2010, Excel sheet

Date	Time	SiteCode	DataValue	VariableCode	SampleType	LabSampleCode	LabMethodID	MethodDescription	OffsetValue	SourceID	QualityControlLevelID	CensorCode	OffsetDescription	OffsetUnitsID	UTCOffset	LocalDateTime
1/5/2010	09:00:00	QAC	0	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	20	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	60	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	80	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	300	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	600	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	1200	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	2000	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00
1/5/2010	09:00:00	QAC	1740	TNH4-DI	1		0	0	Standard	0	#N/A	-9999	nc	Below Surface Water	48	-4	1/5/2010 09:00

Figure 3-3.   Physico-chemical Data Water 2010, with New Formatted Data Values Spreadsheet

3.2.2. Description of Sediment Data File.

The sediment data file contains Fractions and a METADATA spreadsheet. The Fractions spreadsheet contains information about sampling date and time, site location, fraction sizes, and weights. The METADATA spreadsheet has been modified and holds all the Offset Descriptions (Figure 3-4). A new spreadsheet named Variables has been created which holds information for each of the newly created variables. For each fraction size a new spreadsheet has been added to this file. In addition, another spreadsheet named All DataValues has been created. This spreadsheet gathers all the information from each class size and METADATA spreadsheet and creates a CUAHSI formatted DataValue file. This DataValues spreadsheet was copied to a new Excel file and saved as a csv file.

cdate yymmdd10.	pdate yymmdd10.	id $char3.	$char2.	$char2.	rep 1.	et 3.	matrix $char2.	initDW 4.0	finWW 4.0	finDW 4.0	TsedW 15.	TsedW2 5.	H2OV 3.1
collection date	placement date	sample id	long channel id	cross channel id	rep	elapsed time of deployment	Analytical Base (sediment /porosity)	ave cobble tray wt **163** ave small pw sampler wt **53** cobbles+sampler+tray, dry, initial (g)			Total Fraction		
									final wet weight (g)	final dry weight (g)	Gravel Tray_ total sed wt (g)	Gravel Tray_ total sed wt sum of fractions(g)	h2o Volume (ml)
20070424	20070327	OWT	1	1A	1	28	SS	2505				378.5	
20070424	20070327	OWT	1	2A	1	28	SS	2504	3570	3131	680.0		470.0
20070424	20070327	OWT	1	2B	1	28	SS	2499	3525	3132	686.0		424.0
20070424	20070327	OWT	1	3A	1	28	SS	2508				310.1	
20070424	20070327	KRT	2	1A	1	28	SS	2495				574.5	
20070424	20070327	KRT	2	2A	1	28	SS	2504	3586	3211	760.0		406.0
20070424	20070327	KRT	2	2B	1	28	SS	2507	3250	2872	418.0		409.0
20070424	20070327	KRT	2	3A	1	28	SS	2508				242.4	
20070424	20070327	CEC	3	1A	1	28	SS	2499				734.7	
20070424	20070327	CEC	3	2A	1	28	SS	2502	4058	3522	1073.0		567.0

Figure 3-4.   Physico-chemical Data Sediment 2010, with New Formatted Data Values Spreadsheet

5

Figure 3-5 shows the steps taken to load Sources, Sites, Variables and DataValues to the ODM database. It also shows a screenshot of the ODM Data Loader and the ODM Streaming Data Loader.
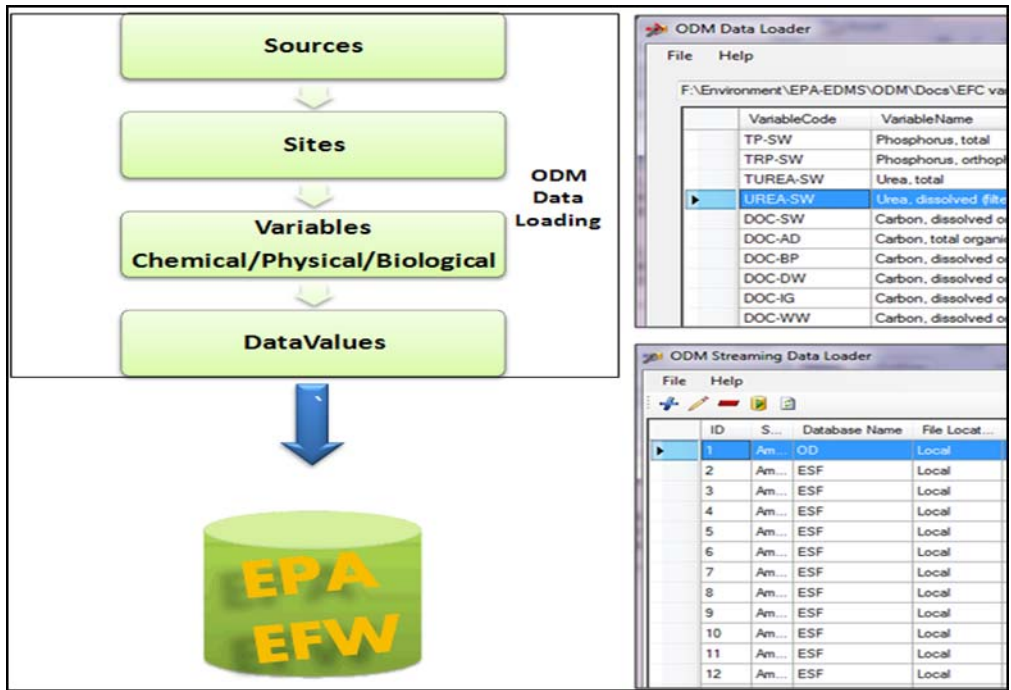
Figure 3-5.   Steps in Data Loading to the ODM
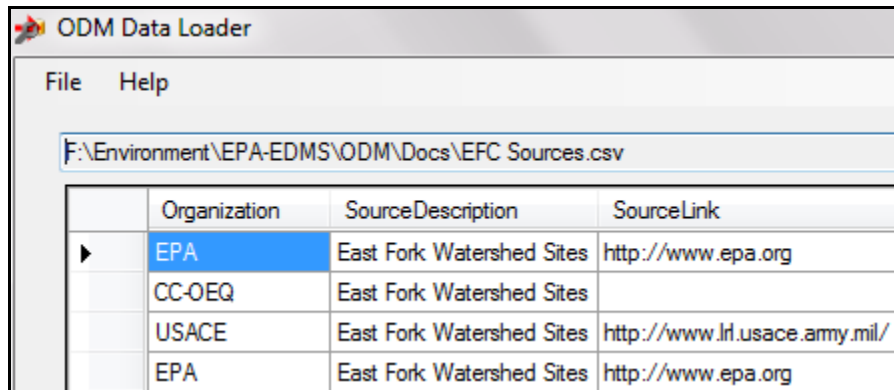
3.2.3. Loading Sites to the ODM

Figure 3-6.   Loading Sites to the ODM

A csv file with all the sampling sites within the EFW was created according to the CUAHSI standard (Figure 3-6). The sampling sites file contains detailed information about site name, code, and location. The site csv file was uploaded to the DMS using the ODM Data Loader.

3.2.4. Loading Sources to the ODM.
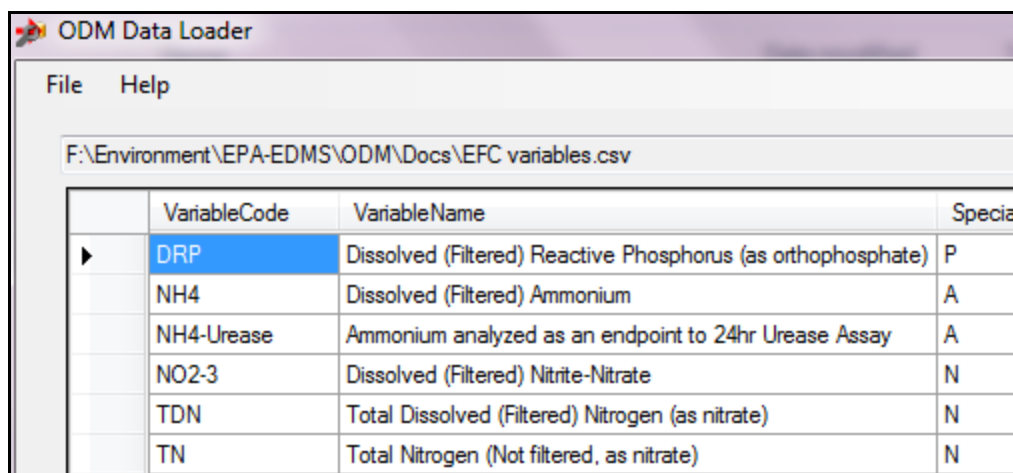
Figure 3-7.   Loading Sources to the ODM

A csv file with information about data sources used in this project was created and loaded to the DMS using the ODM Data Loader (Figure 3-7). This file contains information about the different organizations which mange the different sampling sites within the EFW. It also contains information about a contact person for each of those sampling sites.

3.2.5. Loading Variables to the ODM.

It is sometimes necessary to update the controlled vocabularies. The controlled vocabularies (CVs) in ODM are implemented as tables. For example, the list of valid variable names is stored in the VariableNameCV table (Figure 3-8). To ensure consistent descriptions of user data, the data must use terminology from these CV tables. CUAHSI maintains a master list of CVs that can be synced with an ODM database; however, sometimes it is necessary to add new, project specific terms that aren't in the master CV. This can be done using the tools included with SQL Server.

Figure 3-8.   Loading Variables to the ODM

Before new variables could be loaded to the ODM new vocabularies had to be added to the VariableNameCV table. The new vocabulary was added to the ODM by using SQL statements inside SQL Server.

SQL statement: Insert into VariableNameCV Values ((('Urea, total', 'Total Urea') and Insert into VariableNameCV Values ('Urea, dissolved (filtered)', 'Total Urea')

3.2.6. Loading SampleMediumCV to the ODM.

New vocabularies had to be created and added to the SampleMediumCV table. The new vocabulary was added to the ODM by using SQL statements inside SQL Server.
SQL statement: Insert into SampleMediumCV values('Atmospheric Deposition', 'AD'), Insert into SampleMediumCV values('Intergravel', 'IG'), Insert into SampleMediumCV values('Waste Water', 'WW'), Insert into SampleMediumCV values('Deionized Water', 'DI'), Insert into SampleMediumCV values('Drinking Water', 'DW') and Insert into SampleMediumCV values('Periphyton', 'BP')

3.2.7. Loading VariableNameCV to the ODM.

New vocabularies had to be created and added to the VariableNameCV table. The new vocabulary was added to the ODM by using SQL statements inside SQL Server.
 SQL statement: Insert into VariableNameCV values('Ammonium analyzed as an endpoint to 24hr Urease Assay', 'Ammonium analyzed as an endpoint to 24hr Urease Assay') , Insert into VariableNameCV values('Nitrogen, nitrite (NO_2) + nitrate (NO_3) nitrogen', 'Nitrogen (NO_2) + Nitrate (NO_3) Nitrogen') and Insert into VariableNameCV values('Nitrogen, nitrite (NO_2) nitrogen', 'Nitrate (NO_2) Nitrogen')

3.2.8. Loading SampleTypeCV to the ODM.

New vocabularies had to be created and added to the SampleTypeCV table. The new vocabulary was added to the ODM by using SQL statements inside SQL Server.
SQL statement: Insert into SampleTypeCV values('1', 'Duplicate sample No. 1'), Insert into SampleTypeCV values('2', 'Duplicate sample No. 2'), Insert into SampleTypeCV values('3', 'Duplicate sample No. 3')

3.2.9. Loading Variables to the ODM.

A variable csv file had to be created and formatted according to the CUAHSI standard. All the newly created variables include a unique VariableCode that can be used to specify a variable of interest. The VariableCode starts with the variable abbreviation followed by the sample medium abbreviation to distinguish each variable by its location where it has been collected. For example **TNH4-SW** stands for 'Nitrogen, $NH_3 + NH_4$' and it has been collected as a 'Surface Water' (SW) Sample.

3.2.10. Loading DataValues to the ODM.

The DataValues were created from the 2010_ESF-EFWS_NutrientData.xls file. A new spreadsheet was created and all relevant DataValues Column names added. This spreadsheet pulls values from the Data spreadsheet and the METADATA spreadsheet and writes each value to the correct column in the DataValues spreadsheet. This spreadsheet had to be copied and pasted into a new Excel document. This document was then saved as a csv file (Figure 3-9).

Alternatively, the data can be saved using *Save As CSV* directly from the modified NutrientData.xls file without having to save as a separate Excel file first. The ODM Data Loader was used to load the DataValues to the ODM.

Figure 3-9.   Loading 2010 Nutrient Data Values

3.2.11. Loading Sensor Data from Water-immersed Sensors for Physico-chemical Variables to the ODM.

The sensor water data files were uploaded to the ODM using the ODM Streaming Data Loader (Figure 3-10). All sensor data consider daylight savings time. First a new csv file with all sensor data variables was created. A description for each column in the sensor data files can be found under Appendix C. While loading the sensor data to the ODM database the variables were mapped to the newly created variables in the ODM.

Figure 3-10.   ODM Streaming Data Loader

3.2.12. Loading Sediment Data to the ODM.

The ODM Streaming Data Loader was used to load the sediment data. New vocabularies were added to the VariableNameCV table. The following SQL statement was used inside SQL Server:

Insert into VariableNameCV Values(' Gravel Tray, total sediment weight (g)', ' Gravel Tray, total sediment weight (g)'), Insert into VariableNameCV Values(' Gravel Tray, total sediment weight sum of fractions(g)', ' Gravel Tray, total sediment weight sum of fractions(g)'), Insert into VariableNameCV Values('Gravel Tray , > 2mm (g)', ' Gravel Tray , > 2mm (g)'), Insert into VariableNameCV Values('Gravel Tray, 250um -2 mm (g)', 'Gravel Tray, 250um -2 mm (g)'), sInsert into VariableNameCV Values(' Gravel Tray, 250um - 2mm LOI % (w/w)', 'Gravel Tray, 250um - 2mm LOI % (w/w)'), Insert into VariableNameCV Values(' Gravel Tray, 1.2um - 250um (g)', 'Gravel Tray, 1.2um - 250um (g)') and Insert into VariableNameCV Values(' Gravel Tray, 1.2um - 250um LOI % (w/w)', 'Gravel Tray, 1.2um - 250um LOI % (w/w)')

New sediment variables were created and loaded to the ODM. Sediment Datavalues were loaded to the ODM using a modified version of the ODM Data Loader (Figure 3-11).

Figure 3-11.   Modified Data Loader

3.2.13. Data Validation.

Figure 3-12. New Testing Tab Added to the ODM Data Loader

Three steps were followed to validate the consistency of any data committed to the DMS. In the first step a new tab named Testing in the modified ODM Data Loader was created (Figure 3-12). This tab allows for a quick check and validation of any category of data which has been loaded to the ODM. Under Figure 3-8 the tab 'Units' was queried. In the second step the ODM log files were checked. The ODM data loaders keep a log file which contains information about what type of data has been loaded, the loading date and time, and any errors which occurred during the loading process. The data loaded using the different ODM data loaders were validated after each file had been loaded to the ODM database. All data loaded to the ODM database were quantitatively verified by the user for consistency with the data loaded to the ODM database. The third step included verification that the data values committed to the ODM for a specific location and time were actually in the database, using ODM Tools, a freely downloadable application. This tool enables querying the DMS and verification of the results returned (Figure 3-13) by comparison with the data before loading onto the ODM.

11
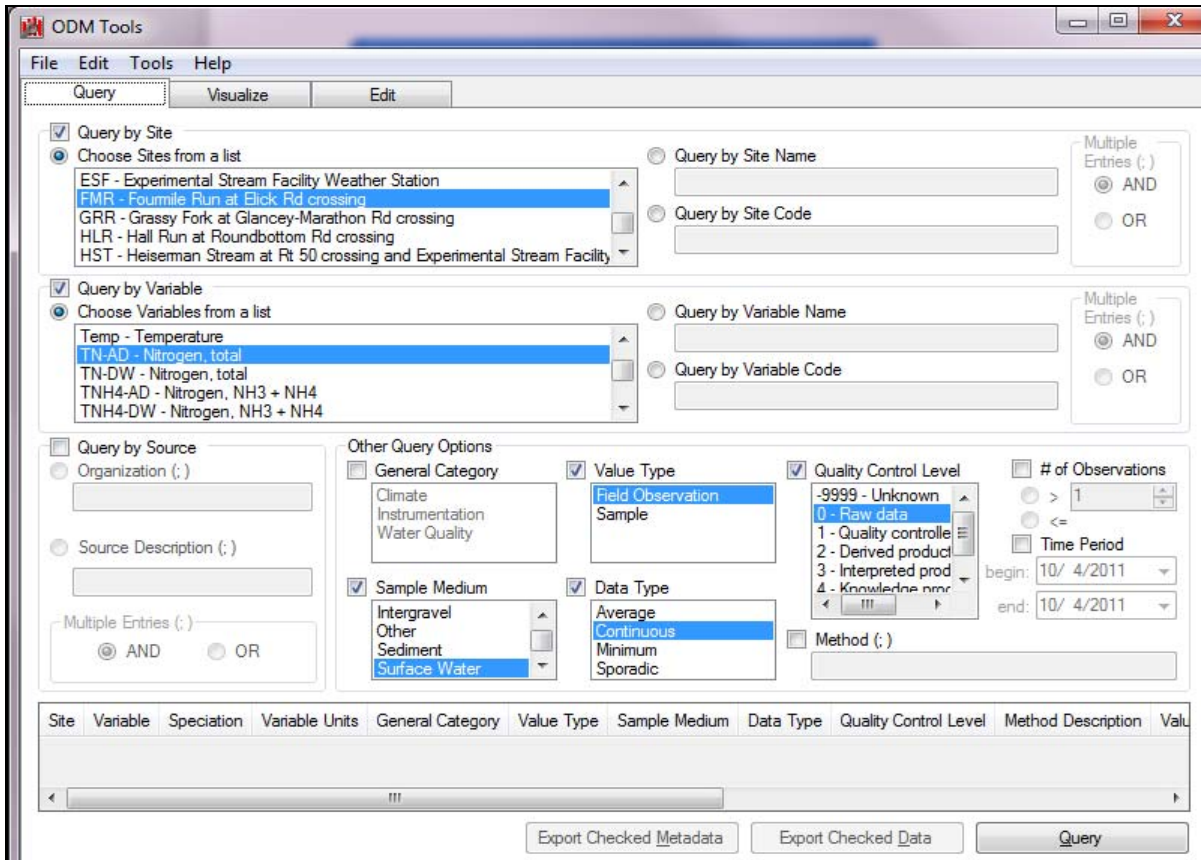
Figure 3-13.   Data Validation Using ODM Tools

3.3   Working with Shepherd's Creek Data

A procedure similar to that used for the EFW data was used to accommodate the SC data into the DMS.In this case only one file type was included, i.e., Excel sheets containing physico-chemicaldata in water grab samples, collected during a 7-year period (2004-2011).

4.0:   DATA PUBLICATION

This section of the report outlines how to publish data through HIS Central and consists of two parts. Part one, *Data Publication with WaterOneFlow*, describes general concepts, requirements and steps to be taken for data publication. Detailed instructions are given in Appendix A. Part two, *Publishing EFW Data*, outlines specific phases towards publishing the EFW data with regard to EPA ORD IT safety requirements.

4.1. Data Publication with WaterOneFlow

WaterOneFlow Web services are used to share data with others online. WaterOneFlow defines a standard set of queries and a standard output format for accessing data, regardless of whether the data are accessed internally from an ODM database, some other database, or even through another website. Additionally, WaterOneFlow provides a layer of security over the database, which makes it less susceptible to hackers than exposing the database itself with public access. Appendix A provides step-by-step instructions on how to publish the data with WaterOneFlow. The main steps include:

1. Create a SQL Server account that the Web service will use to access the database.
2. Install the WaterOneFlow Web service on the computer.
3. Configure the service.
4. Check the result.

Upon the completion of these steps, the service must be registered at HIS Central to make it discoverable online. More than just a listing of WaterOneFlow services, HIS Central performs the following functions:

- Provides detailed information about the service, including contact information, abstract, and areal extent of the sites.
- Supports translation from the variables to an ontology of common hydrologic concepts. This facilitates easy search for variables in the service, especially by those who are not familiar with what the service has to offer.
- Maintains a catalog of sites and variables available in all registered services, enabling fast search for data from multiple data providers.

After the service has been registered, HIS Central needs to *harvest* the data. HIS Central keeps a catalog of all sites and variables, which enables fast searching across all registered services. HIS Central creates this catalog by calling various methods from the WaterOneFlow Web service. This is called data harvesting. When a data harvest is requested by the database developer, an HIS Central administrator has to be notified who then triggers a harvest of data. A trigger request is performed with a click of a button. Once the service is in place, harvesting occurs weekly automatically.

The next step is to map the variables into the CUAHSI HIS ontology. This must be performed for each variable to ensure that variable names comply with CUAHSI ontologies and are,

therefore, recognized. Mapping has to be programmed by the database developer at this time. Finally, HydroDesktop is used to test if data publication has been successful.

4.2. Publishing New Watershed Data

The three phases of publication of a new database include the development/testing, demonstration and production. The team has completed the development/testing and demonstration phases.

4.2.1. DMS Development (and Testing) Phase (Figure 4-1):

1. An ORD-HIS Development Database, an ORD-HIS Test Database, and an ORD-HIS Production Database have been set up by the EPA OSIM inside the EPA firewall. During the development and testing phase, this approach will allow data loaders and application developers and testers three uniform, formatted databases to develop, stage, clean, quality assurance (QA), and test data before it is moved to the production database (which will eventually be accessible to the public).

   DMS Project HIS Database Administrators will have direct access to the three databases (development, test, and production) on the ORD Enterprise SQL Server. The DMS Project HIS Database Administrators will also manage access for other ORD or EPA users (e.g., read only, load data, or modify records).

2. EPA/ORD-users may submit a request for the installation and use of CUAHSI's, open-source, HydroDesktop software, which was recently approved for use within the EPA Agency. Only users inside the EPA firewall can access the EFW and SC databases during the development phase. For demonstration of the EFW and SC databases by the UC contractors, copies on a UC server are used until the production phase is complete.

3. OSIM will also begin setting up an EPA HydroServer that will initially function within EPA's firewall. This HydroServer will live inside the EPA firewall until the DMS team is ready to make it publically accessible. The HydroServer will read data records from the SQL database file that is on the EPA/ORD Enterprise SQL server. The original ESF and SC databases currently reside also on this EPA/ORD Hydroserver at RTP which is accessible from within the EPA/ORD firewall (through a SQL Server account which has permission to view those databases and using ODMTools; point of contact: Dr. Elly P.H.Best).

4.2.2. DMS Demonstration Phase (Figure 4-2):

For demonstration of the EFW and SC databases by University of Cincinnati (UC) contractors, copies on a designated UC server are used until the production phase is complete. These web-services can be accessed by EPA and non-EPA users who have access to the Internet in the HIS Data Service List by the following link: http://hiscentral/cuahsi.org/pub_services.aspx.
By clicking on the title 'EPA- East Fork Watershed in Ohio' the individual service page of the EFW database comes up (http://hiscentral.cuahsi.org/pub_network.aspx?n=264). The data of this database can be accessed via CUAHSI's Hydrodesktop application.
By clicking on the title 'Shepherd Creek Watershed Ohio' the individual service page of the Shepherd Creek database comes up (http://hiscentral.cuahsi.org/pub_network.aspx?n=216).

The data of this database can also be accessed via CUAHSI's Hydrodesktop application. Publication of both databases is visualized in Figure 4-3.

4.2.3. DMS Production Phase (Figure 4-3):

1. Once ORD researchers are ready for the production phase (i.e., data are loaded on the production database and have funding secured), they will begin the Application Deployment Checklist (ADC) or other more recently adopted process to move the EPA HydroServer to the demilitarized zone (DMZ, hosted by the EPA National Computing Center) so it will be fully accessible by HIS Central (unless a different process has been adopted by EPA).

2. ORD and other EPA users with the HydroDesktop client installed will access the internet-based public HIS Central metadata catalog through the Agency Firewall. They will retrieve data records from the EPA HydroServer and from other (public) HydroServers as well. Access to the EPA HydroServer will be through the internal DMZ firewall router, while access to other public HydroServers will be through the Agency Firewall.

3. Public (non-EPA) clients who retrieve pointers to ORD water data from HIS Central queries will access the EPA HydroServer through the Agency internet/DMZ router. The EPA HydroServer will retrieve requested water data records from the internal ORD-HIS database and pass them back to the external client.

4. Offsite (external to the Agency Firewall) access to the ORD-HIS databases will only be possible through the EPA HydroServer, and will be limited to those types of access provided by the HydroServer (i.e., only retrieval of water data records from the ORD-HIS database). Offsite ORD-HIS Database Administrators or other offsite users with requirements to directly access the ORD-HIS databases (e.g., approved external data loaders) will have to use Agency AAA/F5 connection to access the ORD-HIS databases. (An additional 'jump box' may be necessary for offsite direct access to the ORD Enterprise SQL Server and the ORD-HIS databases. This may have to be developed as required).

By placing the HIS database on the ORD Enterprise SQL Server, ORD researchers are relieved of all tasks to maintain the SQL server. ORD researchers (or designated staff) remain the data owner and Database Administrator of the databases. OSIM (or designated staff) manages and maintains the SQL Server itself, including normal server backups. Reliable backup of SQL database files will require arrangement and coordination with the Database Administrator. Additional backup may also be needed and will be determined later.
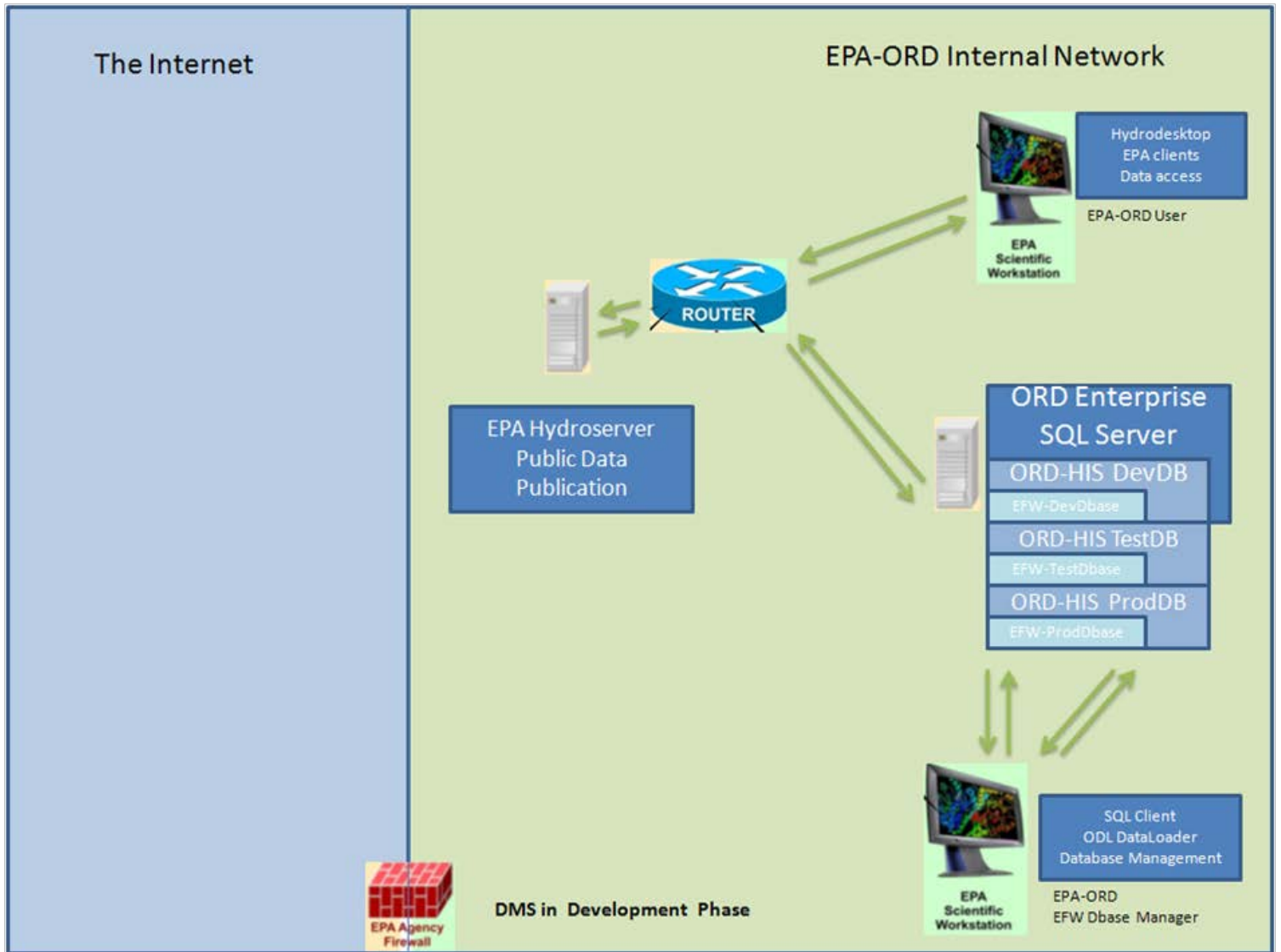
Figure 4-1.   DMS Development Phase
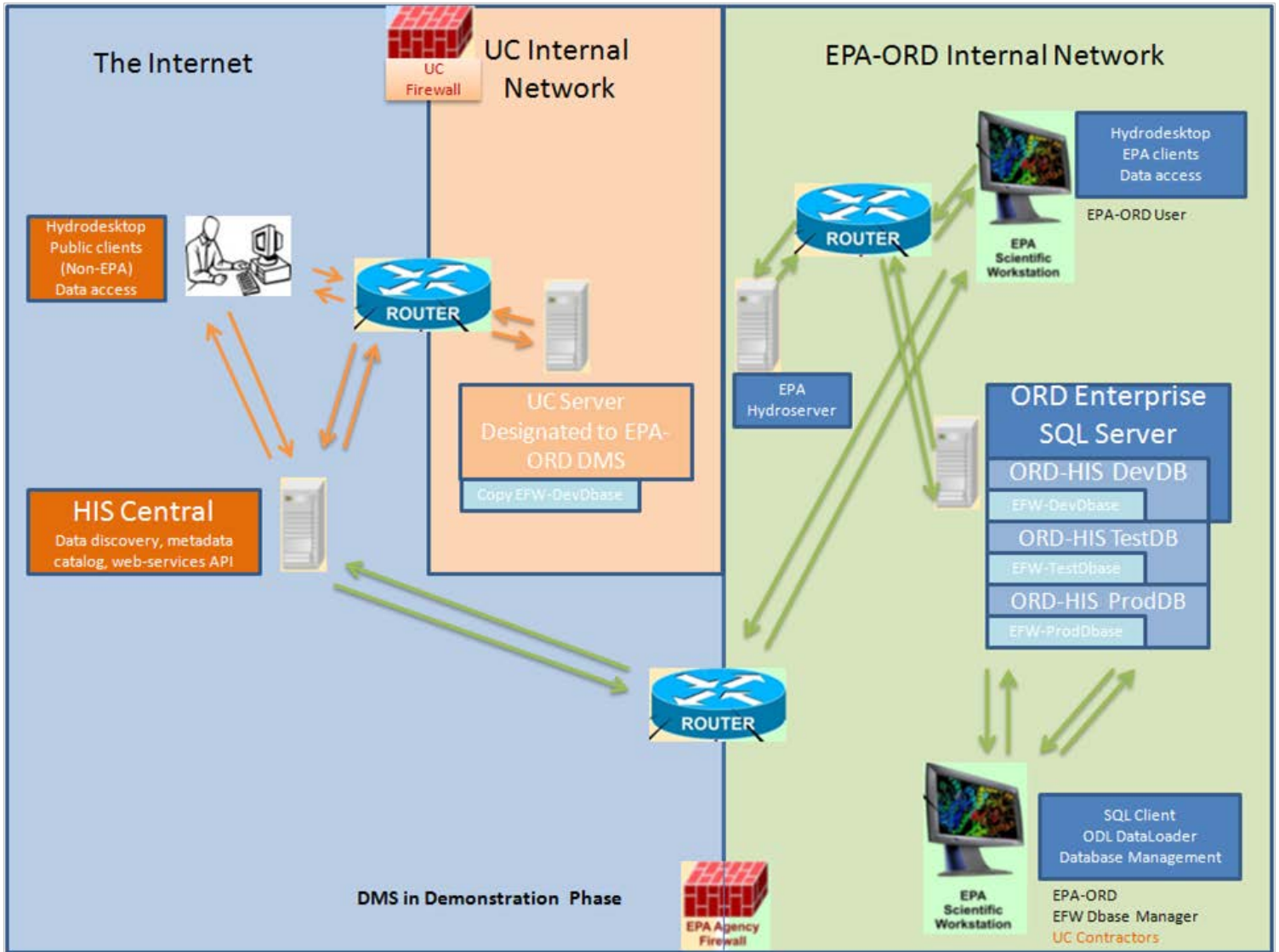
Figure 4-2.   DMS Demonstration Phase

EPA - East Fork Watershed in Ohio

EPA-EFW
EFW

WaterML Service:
http://ly-uc2.ceas1.uc.edu/EFW/cuahsi_1_1.asmx?WSDL
WFS Service:
http://hiscentral.cuahsi.org/WFS/216/cuahsi.wfs?request=getCapabilities
Contact: Amr Safwat

Service Statistics:

Sites:	128	Geographic Extent:
Variables:	55	
Values:	5587751	

Last Harvested on 4/17/2012 8:58:57 AM
(updated weekly, assumed static)

Abstract

Test Data from the EPA East Fork Watershed in Ohio. Data is not quality controlled.

EPA East Fork Watershed in Ohio

Citation

Shepherd Creek Watershed Ohio

Edit Service Details

EPA-SC
Shepherd

WaterML Service:
http://ly-uc2.ceas1.uc.edu/Shepherd/cuahsi_1_1.asmx?WSDL
WFS Service:
http://hiscentral.cuahsi.org/WFS/264/cuahsi.wfs?request=getCapabilities
Contact: Amr Safwat

Service Statistics:

Sites:	6	Geographic Extent:
Variables:	31	
Values:	8312	

Last Harvested on 10/19/2012 1:33:37 PM
(updated weekly, assumed static)

Abstract

Test Data from the Shepherd Creek Watershed in Ohio. Data is quality controlled.

Citation

Test Data from the Shepherd Creek Watershed in Ohio. Data is quality controlled.

Figure 4-3.   Publication DMS in Demonstration Phase

18
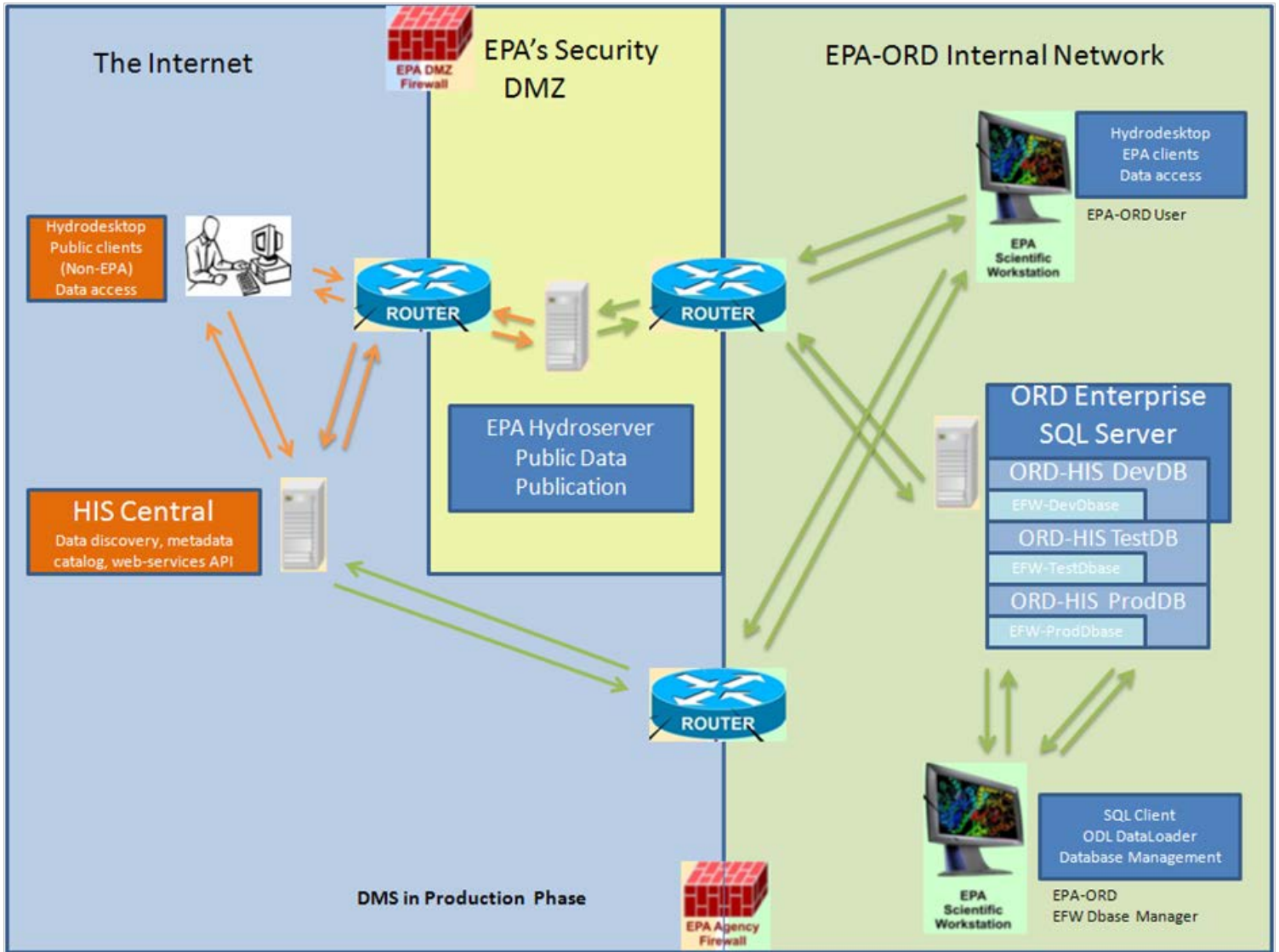
Figure 4-4.   DMS Production Phase

5.0: USING THE DMS: EXCHANGE AND INTEGRATION OF WQ2 INFORMATION

This section describes tools that enable data visualization and analysis. It includes a discussion of HydroDesktop, HydroR and HydroExcel. Details on HydroDesktop and HydroR are contained in Appendix D, and on HydroExcel in Appendix A.

5.1. Using HydroDesktop

HydroDesktop is a desktop application for discovering, accessing, and analyzing time series data from WaterOneFlow services. HydroDesktop can be used to search for hydrologic data, select and download data, visualize time series, label features, delineate watersheds, and explore data.

5.2. Using HydroR

HydroR plug-in is an interface between HydroDesktop and the R statistical software. HydroR makes it easy to provide R access to the data downloaded with HydroDesktop. This interface allows HydroDesktop users access to advanced statistical analysis and plotting capabilities of R.

5.3. Using HydroExcel

HydroExcel is an Excel spreadsheet customized with macros for accessing data from a WaterOneFlow Web service. It is thus possible to extract data from the published Web service from within an Excel spreadsheet.

6.0:   DATA ANALYSIS TO INFORM DECISION-MAKING

6.1. Area of Interest and Problem Identification

In this section a case study is described to illustrate how users may access, explore and retrieve WQ2 data from the DMS, combine them with their own data, conduct analyses and use the results to inform watershed management decision-making.

The East Fork of the Little Miami River (EFLMR) watershed is a large Midwestern watershed, covering approximately 1,295 km^2 (500 mi^2) and discharging into the Little Miami River (LMR), a National Scenic River. The LMR is a tributary of the Ohio River. The headwaters of the EFLMR are located in the rural Counties, Clinton, High, and Brown, and the confluence with the LMR in suburban Clermont County. Dam construction by the U.S. Army Corps of Engineers (USACE) is in 1975 created a 8.74 km^2 (3.38 mi^2) reservoir (Harsha Lake or East Fork Lake). The dam and reservoir provide flood control, recreation opportunities, and a drinking water source for Clermont County. The EFLMR has been listed as an impaired water by the State of Ohio since 2006, and has been designated for Total Maximum Daily Load (TMDL) identification (Collaborative EFLMR Watershed, 2007; Collaborative). The watershed has a mixed agricultural and urban land use, and harbors several wastewater treatment plants (WWTPs) and a drinking water treatment plant (DWTP). Issues of concern for the water quality of this watershed include stormwater, wastewater, and agricultural runoff, which may contribute to increased loading of N and elevated aqueous N concentrations; and drinking water treatability (Nietch et al. 2010). A level of 1.1 mg/L for dissolved inorganic nitrogen (DIN) has been suggested as a water quality standard for the management of rivers and streams <1,300 km^2 (Miltner 2010). In this case study, a procedure is described in which a N-TMDL for the EFLMR is computed using 1.1 mg DIN/L as a target level. The TMDL will be calculated for the pour point of the watershed, where the EFLMR discharges into the LMR. The layout of the watershed is depicted in Figure 6.1.

6.2. Data Management

Data management, sharing, exchangeability, and accessibility were identified as major problems for effective communication and analysis of information by the Collaborative in which EPA/ORD participates. Many WQ2 data are currently collected, analyzed, and stored in similar ways by members of the Collaborative. Therefore, it is expected that a scalable DMS based on CUAHSI protocols for storage and management of georeferenced discrete data pertaining to the EFW collected by EPA will serve as an example and greatly facilitate effective communication and analysis of information in the near future (see sections 2-5 of this report).

6.3. Data Exploration

HydroDesktop was used to explore and identify the geographical area of the EFW, select it, identify the monitoring stations in the watershed, and visualize this information as a geographically referenced map (Figure 6-2).
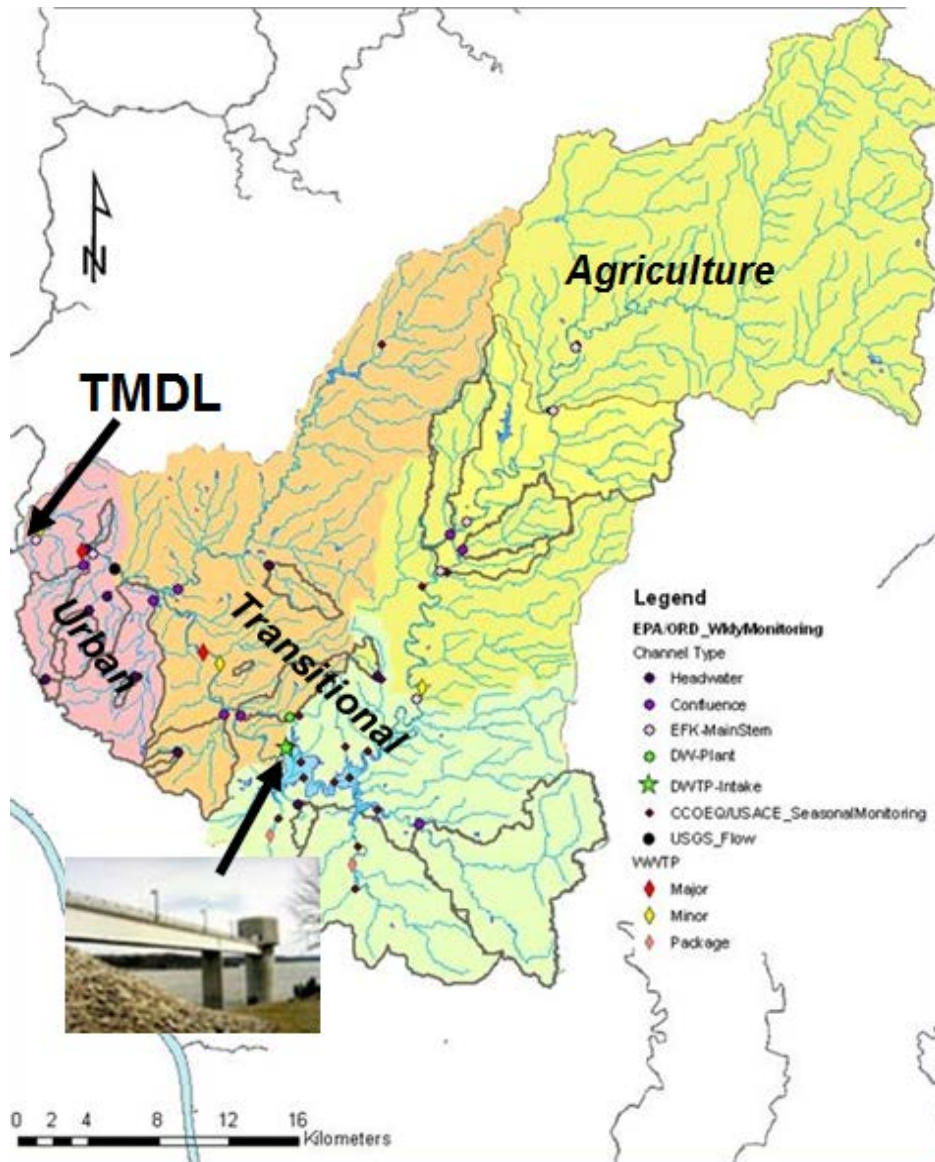
Figure 6-1.   The East Fork of the Little Miami River Watershed, with Point and Non-point Sources for Pollutant Loadings Indicated (from Nietch et al. 2010)

Subsequently, time series and other data available for the monitoring stations were explored, focusing on data and station(s) required for N-TMDL computation, notably nitrogenous compounds and flow at the pour point of the watershed. Data exploration of time series included plotting and computing sample statistics. Flow data from station Perintown, the pour point, collected by the USGS were selected. Within-year seasonality and between hydrological year variation were visualized, via line or scatter plots, and box and whisker plots (Figures 6-3, and 6-4), and summary statistics computed (Figure 6-5). All retrieved data can be exported to other software programs used for further data analysis currently not (yet) provided by HydroDesktop such as Excel.
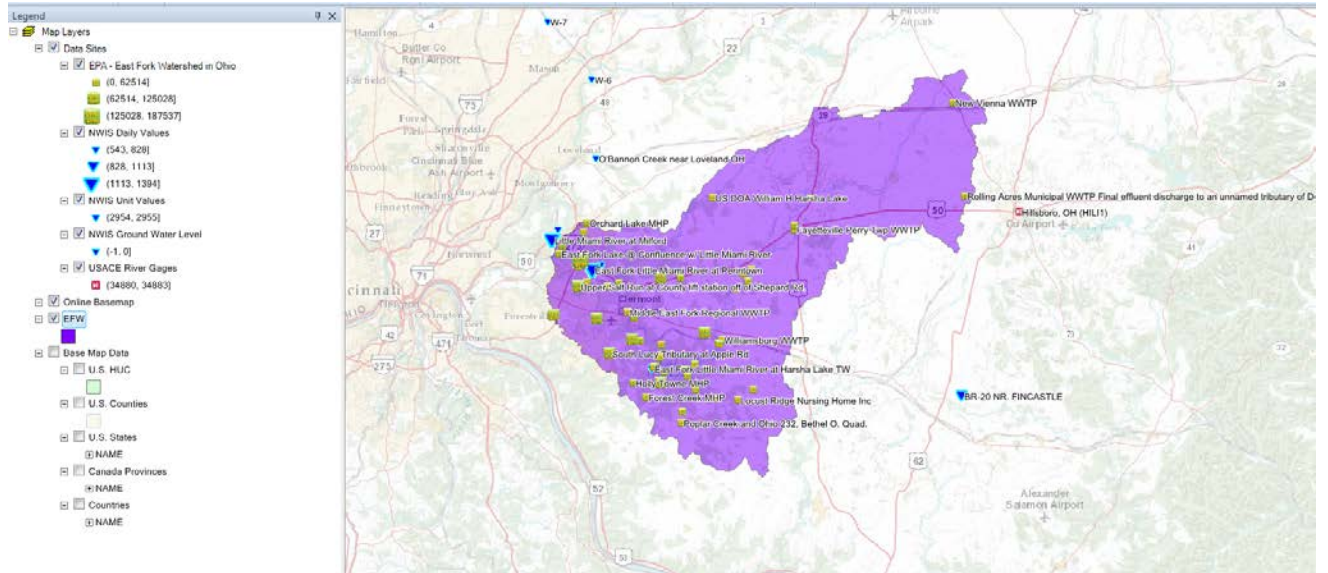
Figure 6-2.   Monitoring Stations for the East Fork of the Little Miami River Watershed
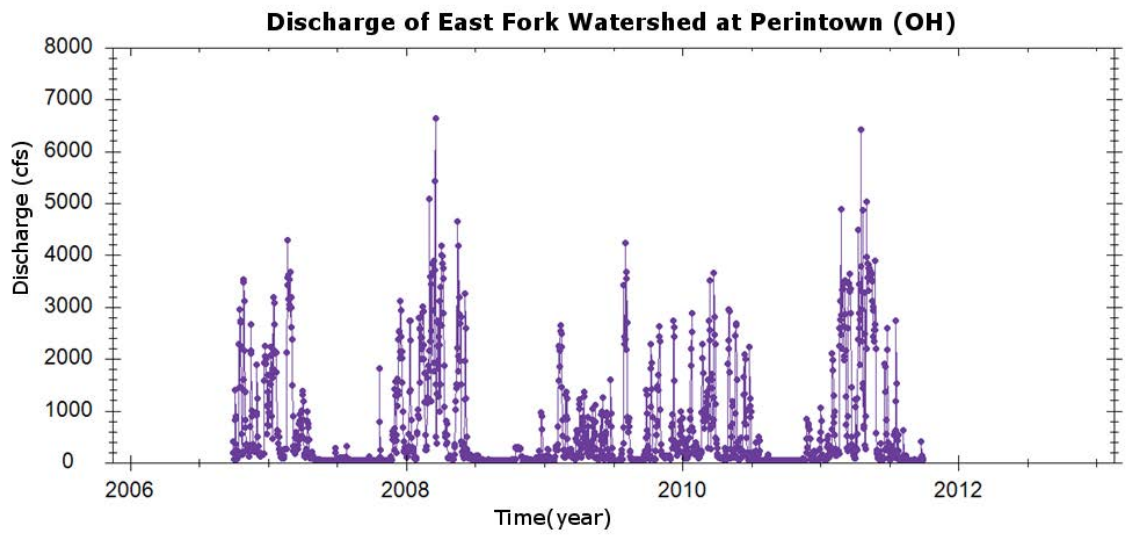
Figure 6-3.   Time Series of the Flow Data at Perintown, OH, Showing Seasonality and Variation by Hydrological Year

23
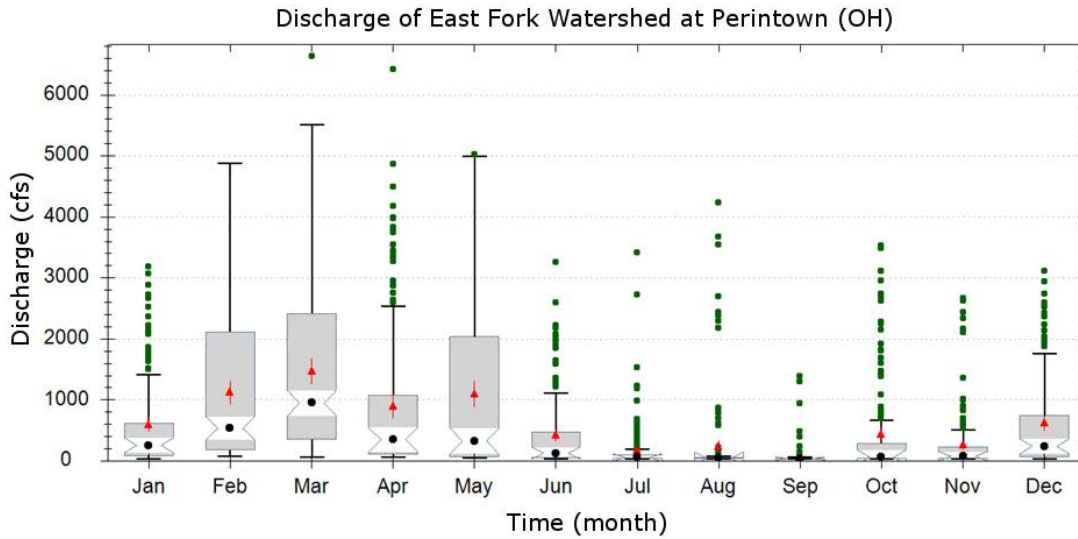
Figure 6-4. **Box and Whisker Plot of the Flow Data at Perintown, OH, Showing a Maximum in March and Minimum in September**

East Fork Little Miami River at Perintown OH	ID 1
# Of Observations	1826
# Of Censored Obs.	0
Arithmetic Mean	617.720153340635
Geometric Mean	199.010893119357
Maximum	6630
Minimum	30
Standard Deviation	978.50988697264
Coefficiant of Variation	1.58406663872119
Percentiles 10%	40
Percentiles 25%	49
Percentiles 50%(median)	140
Percentiles 75%	695
Percentiles 90%	2220

Figure 6-5. **Summary Statistics of the Flow Data at Perintown, OH, Showing Variability in the Data**

6.4. Data Analysis Using Excel

Cumulative flow (in mega gallons per day, MGD) was calculated from measured daily flow (in cubic feet per second, cfs), and desired period of time (in days, d). Cumulative flow over a year is commonly used as the basis for estimation of pollutant loads and/or TMDLs (Figure 6-6). Daily flow is calculated using equation (1).

$$Q_d = \sum_{i=1}^{d} Q_i$$

(1)

where:

Q_d	=	cumulative flow on day d
Q_i	=	flow on day i
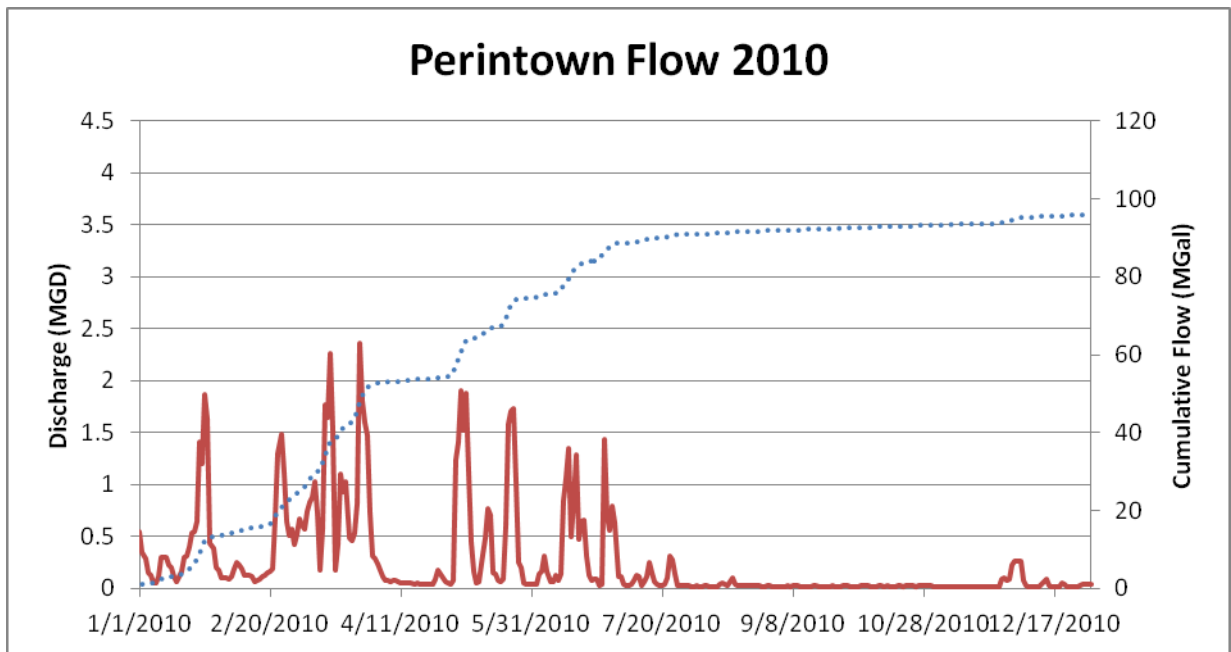
Figure 6-6.   Flow at Perintown, OH, in 2010

In addition to the flow data, data on nitrogenous compounds from station EFW close to Perintown, the pour point, collected by the Collaborative were selected.
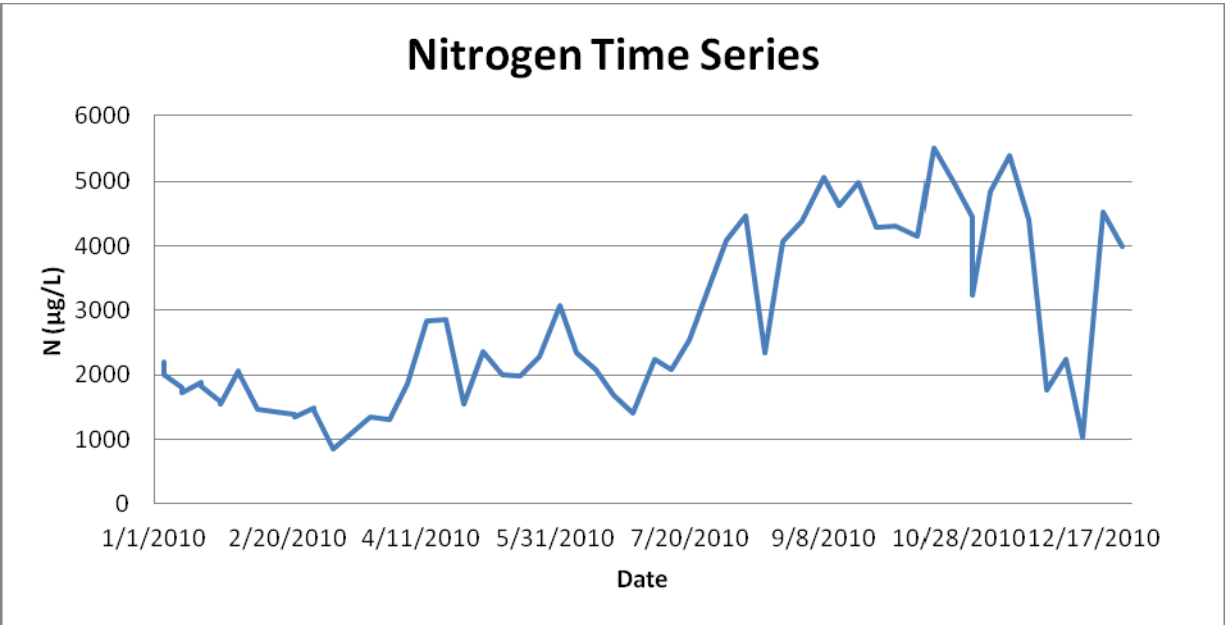
Figure 6-7.   Nitrogen Time Series at Perintown, OH, in 2010

Cumulative load is calculated as follows.

$$L_i = Q_i * C_i \qquad\qquad (2)$$

where:

L_i = load at time i
Q_i = flow at time i
C_i = concentration at time i

Daily cumulative load is then:

$$L_d = \sum_{i=1}^{d} Q_i * C_i * \Delta t \qquad\qquad (3)$$

where:

L_d = cumulative load on day d
Q_i = flow at day i
C_i = concentration at day i
Δt = the dimensionless time step from day d to day $d+1$ at which concentrations were measured

Cumulative target load is calculated as follows.

$$L^C{}_d = \sum_{i=1}^{d} Q_i * C^C * \Delta t \qquad\qquad (4)$$

where:

$L^C{}_d$ = target load at day d
C^C = target concentration selected for this case study, i.e., 1.1 mg/L.

Cumulative nitrogen load and cumulative target load are visualized in Figure 6-8.
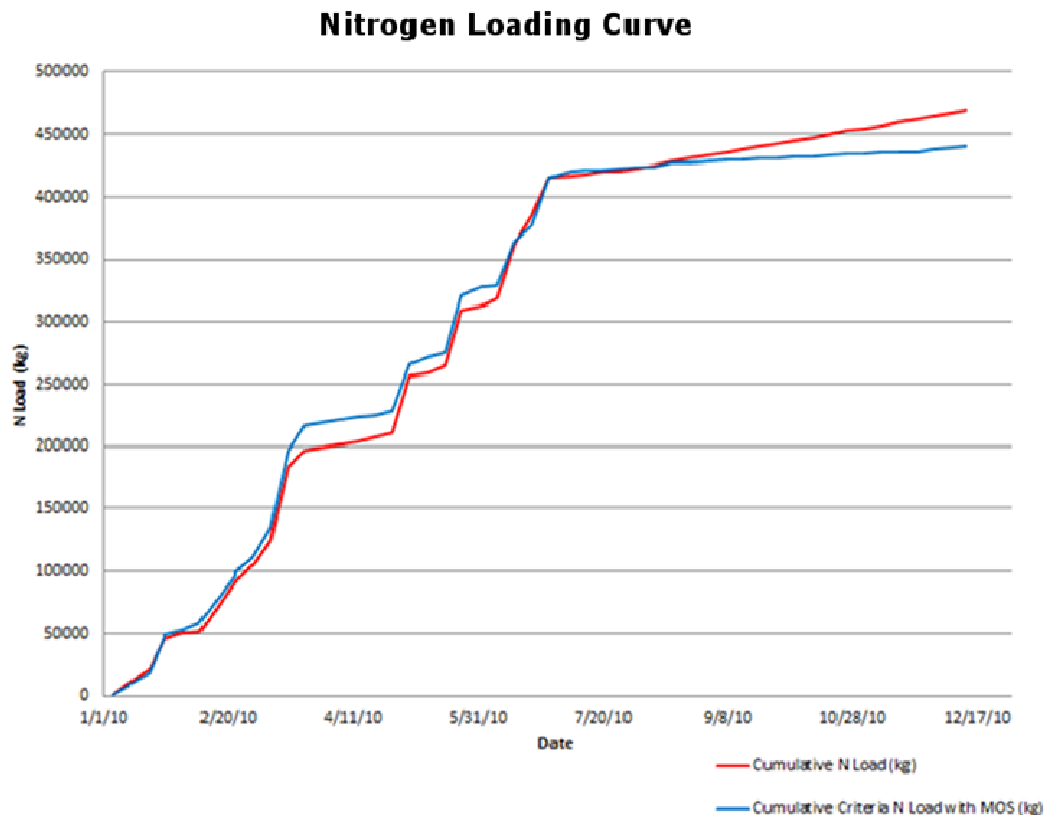
Figure 6-8.   Computed Nitrogen Loading Curve of the East Fork Watershed in 2010

The flow duration curve (FDC) is a plot that shows the percentage of time that flow in a stream is likely to equal or exceed a specified value of interest. Flow duration curves are developed from historic flow data at the site and provide a snapshot of the flow record at that location over a certain period of time.

A duration curve is computed according to the following equation.

$$p(x)=1-P(X \leq x)=1-\sum_{x_i \leq x} P(X=x_i) \qquad (5)$$

where:

p(x)   =   exceedance probability of event x
P(x)   =   probability of event x

The FDC for the pour point of the EFLMR was computed over a period of 5 years from 2005 through 2010. It showed that the flow at this station is at least 100 cfs in 55% of the time (Figure 6-9). Flow duration curves are often segmented into flow regimes to describe varying hydrologic conditions at that site (Cleland, 2003).
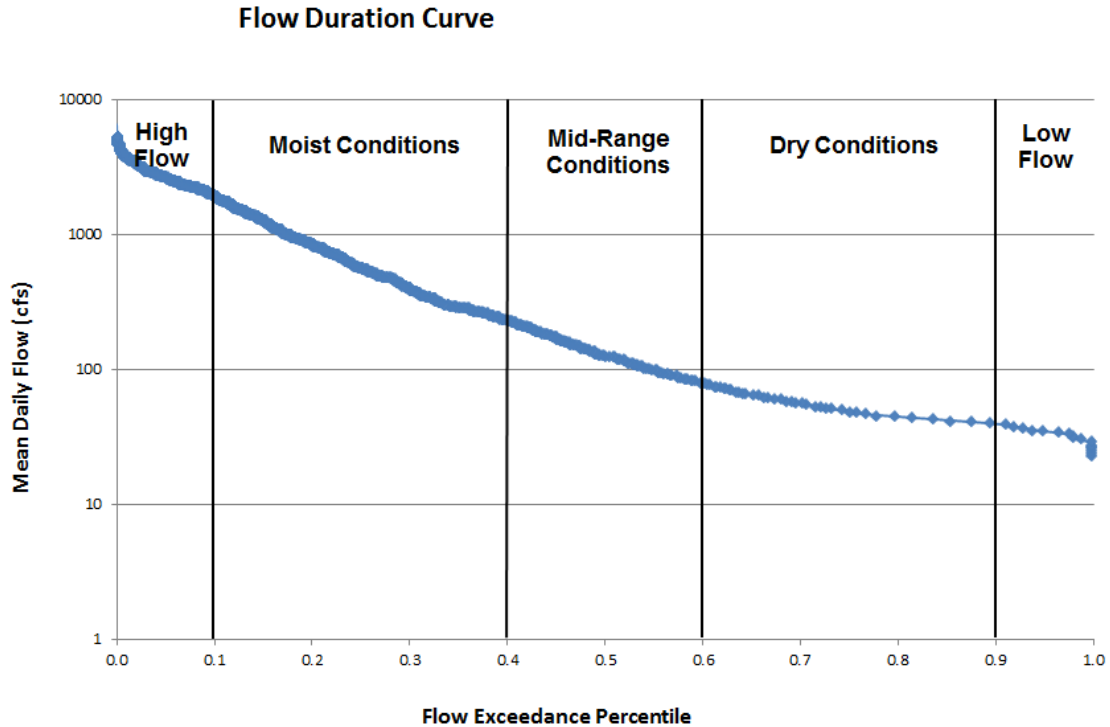
Figure 6-9.   Flow Duration Curve of the East Fork Watershed

Flow duration curves have several applications in the water resources field (Vogel and Fennessey 1995; Johnson et al. 2009), including water quality analysis through load duration curves. Load duration curves are developed from and work similarly to flow duration curves. Instead of addressing flows, however, load duration curves address the likelihood of equaling or exceeding a given pollutant load at a given location. Curves are computed by applying the concept of mass loading which combines water quality and flow information to quantify the pollutant load contributed to a point on a stream by the watershed that lies above it.

The mass loading is then calculated by:

$$L_i = Q_i * C_i \qquad (6)$$

where

L_i	=	mass loading of input at time i in mass/time
Q_i	=	input flow at time i in volume/time
C_i	=	pollutant concentration at time i in mass/volume.

Load duration curves can give insights in various aspects of pollutant loading, such as patterns in loading under various flow conditions, impacts of point versus non-point sources, and effects of best management practices (Cleland 2002; Cleland, 2003; USEPA, 2007). Over the last 10 years load duration curves have been widely used in the calculation of TMDLs.

Using the load duration curves to calculate TMDLs is governed by the following equation

$$TMDL = \sum WLA + \sum LA + MOS \qquad (7)$$

where

WLA   =   the waste load allocation (point sources)
LA   =   the load allocation (non-point sources)
MOS   =   the margin of safety

The Clean Water Act requires that a TMDL includes a margin of safety (MOS) to account for any lack of knowledge concerning the relationship between load and waste load allocations and water quality. According to EPA guidance the MOS may be implicit (i.e., incorporated into the TMDL through conservative assumptions in the analysis) or explicit (i.e., expressed in the TMDL as loadings set aside for the MOS). Commonly a MOS of 5-10% of the TMDL is used. However, a 20% MOS may be used for the load duration curve (LDC) method where a relatively low number of data points is available for the analysis.

For this EFW case, first a target curve was computed from the target concentration (of 1.1 mg N/L) according to equation (7) using a 10% MOS. The resulting curve is the maximum pollutant load that can be experienced at the site, based on previous flow conditions, while still meeting the water quality standard (i.e., the TMDL; Figure 6-10, blue line). Subsequently, the water quality data monitored at the nearby station were converted to loads and plotted in the same graph (Figure 6-10; red dots). This example computation is limited to monitored N values accessible via the DMS, representing non-point sources (*LA*), and, therefore, the estimated N load is probably less than with point-source contributions (*WLA*) included. Because the target curve represents the water quality standard, points falling above the curve are out of compliance, and points below the curve are in compliance. The difference between the observed and target loads are then computed to reveal the overall load reduction required to meet the water quality standard. The required load reduction can be found by inspecting the flow regimes pertaining to the observed and target loads, respectively. The percent load reduction can be calculated by subtracting the average of the observed load from the average of the target load (Figure 6-11). Expressing the load reductions numerically and by flow regime provides more specific information on the circumstances under which TMDLs are exceeded, and may, thus, guide measures and actions required to avoid TMDL exceedance (Table 6-1). To take the MOS into consideration, the target load has to be decreased by 10% for this calculation. In this watershed, the observed N load remained below the target load under high flow conditions in 2010, despite transporting the greatest N load, since ample water was available for dilution. In contrast, under dry as well as low-flow conditions the observed N load greatly exceeded the target load, despite transporting over ten times lower loads per day than at high flow. In this case, therefore, measures alleviating water quality impairments under dry and low-flow conditions may have to be taken. In the current TMDL example computation only N from non-point sources was included. To reduce the N loading from non-point sources, nutrient management measures such as reduced fertilizer rates, conservation tillage, cover crops, and on-site runoff treatment may be considered. Regulatory action recommending lower effluent concentrations for N only pertain to pollutants from point sources.

This case study demonstrated the functionality of the DMS by estimating an example nitrogen (N) TMDL using a Load Duration Curve (LDC) approach. Data were retrieved from the DMS and other databases accessed, explored and visualized through the HIS' application Hydrodesktop. In this case Hydrodesktop accessed the DMS-data via a link to the test-publication of the EFW Database for which CUAHSI provides a codeplex sandbox. When the implementation of the recently developed DMS is approved for use within EPA-ORD and

funding for the DMS is generated to enter into the 'Production Phase' (involving dedicated server testing, operating and maintenance), the publication of the EFW Database in the CUAHSI HIS Services Central Web Service Registry (and possibly other websites) can be accomplished. The latter step would enable access and use by users within and outside EPA/ORD, and is expected to greatly facilitate and accelerate collaboration.
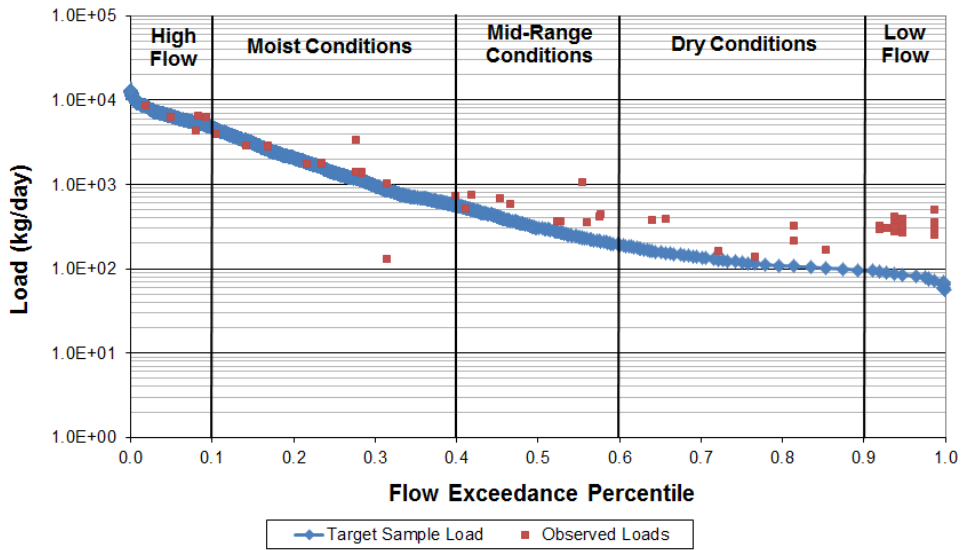
Figure 6-10.   Nitrogen Load Duration Curve of the East Fork Watershed
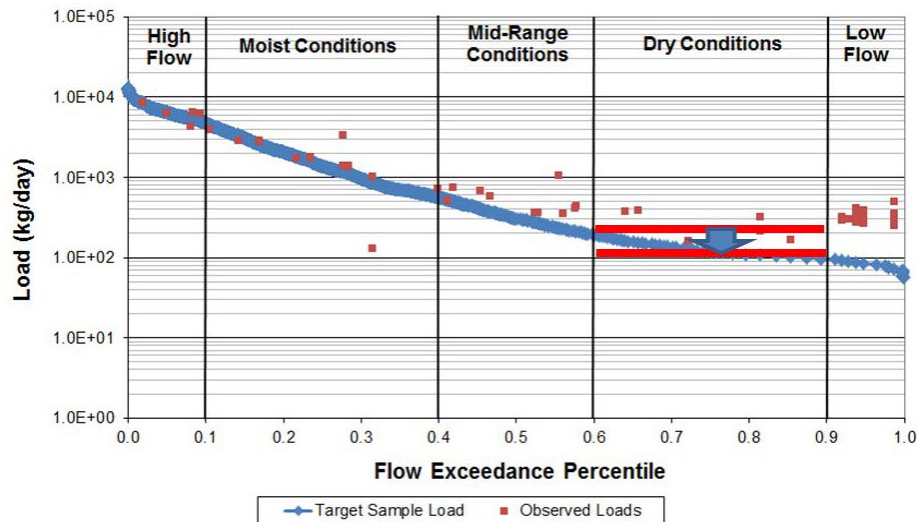
Figure 6-11.   Nitrogen Load Duration Curve of the East Fork Watershed, with Difference Between the Averages of Observed and Target Loads, Respectively, Marked

Table 6-1. **Load Duration Curve Estimated Nitrogen Loads and Overall Reductions Needed to Meet the Proposed Dissolved Inorganic Nitrogen TMDL for the Management of Rivers and Streams <1,300 km^2 (with 10% margin of safety)**

Nitrogen Reductions						
Flow Percentile Range	Hydrologic Condition Class	Number of Samples	Average Observed Load (kg/day)	Average Target Load (kg/day)	Average Target Load w. MOS (kg/day)	% Load Reduction w. MOS =Obs-Target w.MOS)/Obs
0-0.1	High flow	5	6399.39	7608.86	6847.98	0.0
0.1-0.4	Moist conditions	17	1951.91	1964.72	1768.25	9.4
0.4-0.6	Mid-range conditions	10	553.82	369.53	332.58	39.9
0.6-0.9	Dry conditions	7	268.14	141.61	127.45	53.5
0.9-1.0	Low flow	17	335.88	93.27	83.95	75.0

7.0:   CRITICAL GAPS OF CURRENT APPROACHES

The EFW biological monitoring data include data attributes that do not map to attributes within ODM. Because access to the data within CUAHSI HIS is provided by Web services which connect to the ODM database, this presents a number of challenges. The ODM database will not store these attributes. Even if the database were modified to accommodate these attributes, the ODM data loaders would not load these attributes because they have not been programmed to recognize them. Even if the data are somehow successfully loaded into the database, the remaining cyberinfrastructure from the WaterOneFlow Web service to HIS Central to HydroDesktop, would not recognize these attributes and so they would not be communicated to end users. Recommendations for addressing these challenges are described in the next section.

8.0: RECOMMENDATIONS FOR IMPROVING UPON EXISTING APPROACHES

Being an open source system, modifications to CUAHSI HIS software and data models can be made when they fall short of project needs. For example, to address issues with handling biological data as described in the previous section, the following steps could be taken:

1. Using SQL Server, modify the ODM database to include the additional biological attributes.

2. Modify the ODM data loaders to load these attributes. Modifications can be made by downloading the source code from CUAHSI, modifying the code, and compiling the code.

3. While it is not possible to modify the HIS Central application since it is maintained at CUAHSI, users within the EPA firewall can still get access to the data by making direct connections to the ODM database. These connections would allow for the transfer of the additional biological attributes.

A balance must be struck between making these modifications within the DMS team and relying on CUAHSI to perform the work. In some cases, the modifications may prove useful to the broader CUAHSI community outside the scope of the DMS project. CUAHSI provides a number of mechanisms by which the DMS team can contribute to this broader community:

1. When modifying source code, instead of maintaining only a local copy of the code, commit a branch with the code to CUAHSI's source code repository. The two repositories of interest are:

    a. HydroServer http://hydroserver.codeplex.com/

    b. HydroDesktop http://hydrodesktop.codeplex.com/

2. When issues are encountered or to open discussions with the CUAHSI team and community, use the issue tracker and discussion forums located on the websites linked above.

3. When terms must be added to the controlled vocabularies in the project's ODM database, if these terms might be useful to the broader community, submit edits to CUAHSI's master controlled vocabularies at http://his.cuahsi.org/mastercvreg.html.

9.0:   REFERENCES

Cleland, B. 2002. "TMDL Development from the 'Bottom Up' – Part II: Using duration curves to connect the pieces," *In: National TMDL Science and Policy 2002 – WEF Specialty Conference. Water Environment Federation, Phoenix, Arizona.*

Cleland, B. 2003. "TMDL Development from the 'Bottom Up' – Part III: Duration curves and wet-weather assessment," *In: National TMDL Science and Policy 2003 – WEF Specialty Conference, Chicago, Illinois.* p. 27

Horsburgh, J. S., D. G. Tarboton, M. Piasecki, D.R. Maidment, I. Zaslavsky, D. Valentine, and T. Whitenack. 2009. "An Integrated System for Publishing Environmental Observations Data," *Environmental Modeling and. Software*, 24, 879-888.

Horsburgh, J. S., D. G. Tarboton, D.R. Maidment, and I. Zaslavsky. 2008. "A Relational Model for Environmental and Water Resources Data," *Water Resources Research*, 44, W05406, 12 p.

Johnson, S.L., T. Whiteaker, and D.R. Maidment. 2009. "A Tool for Automated Load Duration Curve Creation," *Journal of the American Water Resources Association,* vol. 45, No 3: 654-663.

Miltner, R.J. 2010. A Method and Rationale for Deriving Nutrient Criteria for Small Rivers and Streams in Ohio. *Environmental Management.* DOI 10.1007/s00267-010-9439-9.

Nietch, C., M. Elovitz, E. Heiser, H. Thurston, L. Underwood, H. Lubbers, D. Macke, E. Best, P. Braasch, R. McClatchey, D. Brown, M. Heberling. 2010. "Linking sources to stress dynamics in drinking and recreational waters in a mixed-use watershed in Southwestern Ohio with a multi-agency cooperative. *Society of Environmental Toxicology and Chemistry North American 31st Annual Meeting, Portland, OR, November 7-11, 2010.*

Tarboton, D.G., J.S. Horsburgh, and D.R. Maidment. 2008. "CUAHSI Community Observations Data Model (ODM) Version 1.1 Design Specifications," CUAHSI.

USEPA (U.S. Environmental Protection Agency). 2007. "An approach for using load duration curves in developing TMDLs," *U.S. Environmental Protection Agency: Office of Wetlands, Oceans, & Watersheds, Washington, D.C.,* 68 p.

Vogel, R.M. and N.M. Fennessey, 1995. "Flow Duration Curves II: A Review of Applications in Water Resources Planning," *Water Resources Bulletin* 31(6):1029–1039.

This page intentionally left blank.

APPENDIX A:

Workshop: Publishing Data with the CUAHSI Hydrologic Information System

September 7, 2011

by:

Dr. Tim Whiteaker (twhit@mail.utexas.edu)
Center for Research in Water Resources
The University of Texas at Austin

Distribution

Funding and Acknowledgements

Table of Contents

Introduction

This document provides steps to complete the hands-on training portion of a workshop that teaches how to publish water observations data using the CUAHSI Hydrologic Information System (HIS). This involves loading data into an Observations Data Model (ODM) database, exposing the data in a secure and standard way online via WaterOneFlow Web services, and making the data discoverable by registering the Web service with HIS Central.

For background information on CUAHSI or HIS, please refer to presentation materials provided at the workshop, or the HIS website at http://his.cuahsi.org.

Why publish data online?
Sometimes people ask what the motivation is for using HIS to publish data. When the HIS team interacts with HIS users, here are the most common reasons those users give for why they want to publish their data:
- Academics
  - Recognition of work
  - Data publication is mandated by the funding agency
  - To support science in the US and promote collaboration
- Agencies
  - Standardizing data access (both internally and externally)
  - Time savings in developing a publication system
  - Public benefit with publication
  - Return on investment – people can get the data themselves without requiring a "middle-man"
  - Get all the state or regional data "together"

Additionally, HIS just makes it easy for users to discover and access data. This is traditionally a pretty big time sink for users, so let's be good citizens and make their lives (and ours) easier!

If you have other reasons for publishing data, please let us know! Fulfilling your needs is a primary driver in the future development of HIS.

Goals of the Workshop

This workshop seeks to introduce you to the **HIS data publication process**. As there are numerous avenues for publishing data, which often depend on a given user's available software and system setup, this workshop does not seek to teach every technique that can be used for publishing data. Rather, the workshop communicates the basic concepts of data publication, which can then be applied by you to fit your specific needs or environment. After completing the workshop, you should be able to:
- Understand what kind of data can be stored in an Observations Data Model database
- Translate your observations data and metadata into terminology used by the Observations Data Model
- Load your data into an Observations Data Model database
- Publish your data with a WaterOneFlow Web service
- Register your Web service with HIS Central so that others can discover it
- Access your data in a number of ways using HIS software:

- o   Direct database connection with ODM Tools
- o   Direct Web service access with HydroExcel
- o   Discover and access with HydroDesktop
- Learn more about HIS using the HIS website at http://his.cuahsi.org

Workshop Requirements

Computers and sample data were prepared ahead of time for the workshop. However, all HIS software is free, and so the configuration below can also be applied to your own computer if you have licenses for the commercial software used, such as the operating system. The system outlined below closely resembles the HydroServer system described at http://his.cuahsi.org/hydroserver.html, but without website and spatial data hosting capabilities.

> **Note**
> The instructions in this manual are written assuming your computer is configured like the one described below.

Hardware:
- PC, 4 GB RAM, 500 GB hard disk, 2.7 GHz processor
- Network setup allowing external access to Web services installed on local PC

Software:
- Windows 7
  - o   Microsoft Internet Information Services (IIS) 7 — comes with Windows 7, but may need to be enabled.
    - ▪   Enable ASP.NET.
- .NET Framework 2.0 SP2, 3.5 SP1 (free)
- Microsoft SQL Server 2008 R2 (commercial) or SQL Server 2008 R2 Express (free)
  - o   Be sure to install the version with SQL Management Studio (aka Database and Management Tools)
  - o   Install with these options:
    - ▪   Install with Mixed Mode Authentication.
    - ▪   Specify the "sa" (system administrator) password that you will remember.
- HIS software (free)
  - o   ODM Data Loader - http://his.cuahsi.org/odmdataloader.html
  - o   ODM Streaming Data Loader - http://his.cuahsi.org/odmsdl.html
  - o   ODM Tools - http://his.cuahsi.org/odmtools.html
- Additional software (not required for HydroServer)
  - o   Microsoft Office 2010 (32-bit version)
  - o   Google Earth (free) - http://earth.google.com/download-earth.html
  - o   HydroObjects (free) - http://his.cuahsi.org/hydroobjects.html
  - o   HydroExcel (free) - http://his.cuahsi.org/hydroexcel.html
  - o   HydroDesktop (free) - http://his.cuahsi.org/hydrodesktop.html

Data (located on your Desktop in a folder called Workshop):

- Raw data files of water quality time series
- Metadata text file describing the data
- Solution files, for reference or for use if workshop steps cannot be completed successfully
- Image files that will be associated with your data

User (You):
- Basic knowledge of how to operate a computer and use the internet
- Very basic notions of database concepts such as the terms "table" and "field"
- Rudimentary understanding of hydrology and hydrologic data

About the Workshop Data

In this workshop, you will publish time series of water quality data measured for the Lake Champlain Long-term Water Quality and Biological Monitoring Project. The data include measurements of nitrogen, phosphorus, temperature, total suspended solids, and chlorophyll a, taken from 1992 to 2007. More information about this project can be found at http://www.anr.state.vt.us/dec/waterq/lakes/htm/lp_longterm.htm.

Each workshop computer has been installed with "raw" data files of the water quality observations described above. Some modifications have already been made to the raw files to facilitate data loading. We wanted you to get a sense of how to transform data, without overburdening you with easy-but-tedious data operations. The raw data files include site locations, time series of water quality observations, and metadata. These files are located on your Desktop in **Workshop\RawData**. During the workshop, you will transform the raw data files so that they are in a form that can be loaded into an Observations Data Model database. As a contingency plan in case you are unable to complete the data transformation process, transformed files have been generated for you and are located in **Workshop\SolutionFiles\TransformedData**.

Review of CUAHSI-HIS Data Publication Tools

The CUAHSI Hydrologic Information System provides Web services, tools, standards and procedures that enhance access to more and better data for hydrologic analysis. HIS software is free and available on the HIS website at http://his.cuahsi.org. A variety of HIS software applications have been built to serve several types of users and scenarios, from data users to data publishers to educators and developers.

HydroServer
In this workshop, you'll be playing the role of the data publisher. This means you have some hydrologic observations data that you've collected, and you'd like to publish that data on the Web in a standard way so that others can easily access and use it. To facilitate data publication, HIS offers **HydroServer** (http://his.cuahsi.org/hydroserver.html), which is really just a bundle of HIS software designed for data publication that operates on a Windows computer.

HydroServer includes software for publishing observations data with a WaterOneFlow Web service, but it also includes a website and supporting components for geospatial and temporal data visualization. These website components are not required for data publication and are not covered in this workshop. The focus of this workshop is on storing observations data in an Observations Data Model database and publishing the data with a WaterOneFlow Web service.

Observations Data Model

The **Observations Data Model (ODM)** (http://his.cuahsi.org/odmdatabases.html) is a data model for the storage and retrieval of hydrologic observations in a relational database. An ODM database stores data and sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used. A relational database format is used to provide querying capability to allow data retrieval supporting diverse analyses. To learn all the details of ODM, read the design specifications document on the website linked above.

Data can be loaded into an ODM database using a number of tools, including free HIS software. For loading static data files for what is generally a one-time process, the free **ODM Data Loader** is used. These data files are usually the result of a study or project that has been completed and will not need periodic updating. The ODM Data Loader is optimized for data with one data value per row. For data that are continuously updated, such as data streaming in from sensors in the field, use the free **ODM Streaming Data Loader**. The Streaming Data Loader is optimized for data with one time step per row. For more complex data loading tasks, SQL Server Integration Services is one of many software packages up to the task. However, these software packages are typically not free and require significant training to learn how to use them. For this workshop, you'll gain experience with both ODM data loaders.

You'll load most of the data in this working using the ODM Data Loader. The ODM Data Loader reads input files that are formatted much like the tables in ODM. For example, if you want to load site locations into ODM, you could prepare a spreadsheet called "sites.xls" with column headings that use (roughly) the same names as fields from the Sites table in ODM (names are not case sensitive). In some cases, you can load data for more than one ODM table from a single input file by simply appending additional columns to the data in the file. This prevents you from having to create an input file for every table in ODM, which would be quite tedious since relational databases tend to have many associations across many tables. A document describing the input format required by the ODM Data Loader can be found at http://his.cuahsi.org/odmdataloader.html.

Once data are loaded into an ODM database, you can examine the data using the free **ODM Tools**. Or, if you have knowledge of SQL, you can write your own queries within SQL Management Studio. If the data look good, you can publish the data with a WaterOneFlow Web service.

WaterOneFlow
A challenge in querying and interpreting data from disparate data sources is that each data source not only has its own method for asking for data, but also its own format for delivering the requested data to the user. **WaterOneFlow** overcomes this by providing a single query interface and a standard output format called WaterML, which is an XML language for the communication of water data. WaterOneFlow is a Web service, which facilitates automated and programmatic access to the data. This is an advancement beyond simply publishing data on a web page, which can require complicated and often error-prone screen scraping and parsing.

A WaterOneFlow Web service is available that hooks directly into an ODM database to publish data from that database. However, WaterOneFlow Web services can also be written to support internal data formats other than ODM. This means that no matter what data storage mechanism you choose to use, you can still publish your data in a standard way with WaterOneFlow.

HIS Central
Once your data are published, there's still the issue of data discovery. How do people learn about your data? That's where HIS Central comes in. **HIS Central** is a website maintained by the CUAHSI HIS team where you can register your WaterOneFlow Web service. Your service then becomes discoverable along with dozens of other Web services already registered with the system, including services for EPA

and the USGS National Water Information System. This makes HIS Central the largest single catalog of the nation's water data.

HydroExcel and HydroDesktop
Now it's time to briefly play the role of the data user. Once data are published with HIS, how do people access the data? To help data users get started with HIS, several free applications or application extensions are available on the HIS website, geared towards application environments most commonly used by hydrologists such as Microsoft Excel. **HydroExcel** is an Excel spreadsheet customized with macros for accessing data from a WaterOneFlow Web service. **HydroDesktop** is a free and open source Geographic Information System (GIS) designed from the ground up to work with HIS. In the workshop, you'll use both HydroExcel and HydroDesktop to verify that you have successfully published your data with WaterOneFlow.

Workshop Outline

The workshop begins with presentations and demonstrations by the HIS team to familiarize the audience with HIS. Contact the workshop administrator to check for availability of these materials. The hands-on training portion of the workshop leads the audience through the data publication process with these key steps:

1. Translate raw data for loading into ODM.
2. Load data into an ODM database.
3. Expose database content online via WaterOneFlow Web service.
4. Register Web service with HIS Central to enable data discovery by external users.

Translating and Loading Data into ODM

To load data into ODM, you'll be using a tool called the ODM Data Loader. The ODM Data Loader loads data from comma delimited files (.csv) or Microsoft Excel 2003 files (.xls) that have a one row header that uses ODM field names in the header, followed by the data in subsequent rows. When loading data from Excel, the data should be located in a worksheet that has the same name as the file. More about these data formats can be found in the documentation for the data loader available at http://his.cuahsi.org/odmdataloader.html. Generally, the fields in the input files conform to the table structure of ODM, with some flexibility for specifying alternative information for database generated IDs. You'll see how this works during the workshop.

Most likely, your data are not exactly in the same format as what the ODM Data Loader is expecting. For example, instead of using the ODM terminology "SiteCode", you may call the unique ID for each of your observation sites a "StationID". Also, you may need to do some leg work to look up information such as the horizontal datum associated with the latitude and longitude coordinates of your site, which is one of the pieces of information that ODM requires. Translating your data and metadata into ODM terminology may seem a bit tedious, but this exercise is actually very valuable as it helps you to fully understand ODM as well as your own data, and in the end you will have a database that richly describes your data.

In the interest of time, much of the data translation work has been performed for you. However, some items still remain untranslated and are just dying to have a talented hydro hero like you perform the transformation and save the day!

Installing a Blank ODM Database

An understanding of your own data as well as the Observations Data Model is essential before attempting to transform your raw data into inputs for the ODM Data Loader. To help you in this process, ODM includes some predefined terms called **controlled vocabularies** that you can choose from when populating its tables. Let's grab an ODM database from the HIS website and see some of these terms for ourselves.

To attach a blank ODM database:
1. In a web browser, navigate to http://his.cuahsi.org.
2. Under **Quick Links** on the right, click **ODM Database**.
3. Click the link to download the **ODM 1.1 Blank SQL Server Schema Database**.
4. Unzip the contents of the downloaded file into the **SQL Server data folder**, e.g., C:\Program Files\Microsoft SQL Server\MSSQL10_50.MSSQLSERVER\MSSQL\DATA. The download includes both a database file (.mdf) and a log file (.ldf) that tracks transactions made to the database.

You'll now attach this database to SQL Server, and eventually you will load the workshop data into this database.

Tip
SQL Server is the database software installed on the workshop computers. "Attaching a database" to SQL Server basically means letting SQL Server know about your database so that its software can work with it.

5. Start **SQL Server Management Studio**.
6. Make sure the **Server name** is **(local)**.

7. Click **Connect** to log into SQL Server.

8. In the Object Explorer on the top left, right-click **Databases** and click **Attach** (Figure 0-1).
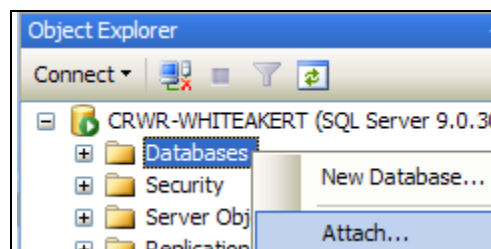
Figure 0-1 SQL Server Management Studio is used to attach databases to SQL Server

9. In the Attach Databases dialog that opens, click **Add**.
10. Navigate to the folder where you saved your database, select **OD.mdf**, and click **OK**.
11. Change the "Attach As" name to **MyWaterData** (Figure 0-2), and click **OK**. (You can change the name to whatever you want, but let's all use MyWaterData as the name for this workshop exercise.)
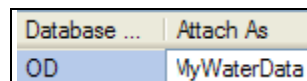
Figure 0-2 You can assign a database name of your choice in SQL Server

12. Click OK if prompted about full-text catalogs.
13. In the Object Explorer, click the plus sign to expand **Databases**. You should now see your MyWaterData database.

Now that the database is attached, let's look its tables.

To explore the database you just attached:
1. In the Object Explorer, expand **MyWaterData**, and then **Tables**, to see a list of tables in this blank ODM database.
2. Open the **DataValues** table by right-clicking it and then clicking **Select Top 1000 Rows**. This table stores time series values. Each row stores a single datetime, a single value, and metadata about that value. The table is currently blank, but you'll fix that later on!
3. Notice that there is a field in this table called **SiteID**. Rather than repeat the latitude, longitude, and other site information with every row in the DataValues table related to a given site, ODM

keeps the database compact by using the SiteID to locate a matching row in the Sites table where the site details are stored. These **relationships** between tables are used extensively throughout ODM, leveraging the power of a relational database.

4. Click the X near the top right of the query that was opened to close the query and the table. Be sure not to click the X in the blue title bar for the application, or you will close Management Studio!

5. Open the **Sites** table to see the information describing each site. **Close** the table when you are finished looking at it.

So far, the tables that you've been looking at are empty. Next, you'll open some tables that have already been populated with values that will be useful as you load data.

6. Open the **VariableNameCV** table. The letters "CV" at the end of the table name indicate that this is a controlled vocabulary table. Only terms from this table are used to describe variable names. This helps to standardize the terminology used to describe data across multiple ODM databases (Figure 0-3).

Term	Definition
Momentum flux	Momentum flux
N, albuminoid	Albuminoid Nitrogen
Net heat flux	Outgoing rate of heat energy transfer minus the inco...
Nickel	Nickel (Ni)
Nitrogen, Dissolved Inorganic	Dissolved inorganic nitrogen
Nitrogen, Dissolved Organic	Dissolved Organic Nitrogen
Nitrogen, gas	Gaseous Nitrogen (N2)

Figure 0-3 The VariableNameCV table is a controlled vocabulary of terms to use when naming variables

Tip
A level of data integrity is enforced through the use of controlled vocabularies (CVs) within ODM. If a field uses a CV, then only terms from that CV can be entered into that field. This way, the database uses consistent terminology. If there is a CV term that you need which is not already in ODM, you can add it. Just remember to add the term to the CV table first, and then load your data.

7. Open the **SampleMediumCV** table. This table has terms used to describe the sample medium in which a measurement applies.

8. Continue opening and browsing the tables of ODM. When you are ready to move on, **close the tables** and **minimize SQL Server Management Studio**. You can leave Management Studio open as we'll use it again later.

The controlled vocabularies were originally conceived by the HIS team and are updated by users like you who submit change requests or additions to the CV website at http://his.cuahsi.org/mastercvreg/cv11.aspx. As this list is updated, the CVs in your database may become out of date. The ODM Tools makes updating your CVs a snap. You'll use the ODM Tools again later to analyze the data you load, but now let's use the tools just to update the CVs in your database.

The ODM Tools software is a free download from the HIS website at http://his.cuahsi.org/odmtools.html. The tools have already been installed on the workshop computers.

To update controlled vocabularies using the ODM Tools:
1. Start the **ODM Tools**. When you run the ODM Tools for the first time, it will prompt you for a database connection.
2. Input these values into the New Database Connection dialog and then click **Save Changes**.
    a. Server Address: (local)
    b. Database Name: MyWaterData
    c. Server User ID: sa
    d. Server Password: [enter the password for the server user ID here]
3. Dismiss the message indicating that the connection was successful.

Tip

The sa account is an administrator account for SQL Server. It has all the privileges we need for this workshop and more. For security in your production environment, you may want to establish a SQL Server login with lower permissions and call it ODMToolsUser, for example, so that your users do not need to know the administrator login and password. Steps for creating a login are provided later in this manual.

When the tools open, the interface looks a bit empty. That's because you haven't loaded data into the database yet. However, we can still update the controlled vocabularies.
4. Click **Tools | CV Update**.

A dialog opens showing you a comparison between the master controlled vocabulary list maintained by CUAHSI (on the left) and your local CV (on the right). You can change the CV that you are looking at using the drop down box in the top left corner of the dialog.
5. Change to the **Sample Medium CV**.

As you can see, (at the time this document was written) the master list has been updated since the last time the ODM blank database was created for the HIS website.
6. Click **Update Local CV**.

Your local CV now indicates that terms have been added but not committed.
7. Click **Apply**. This commits the changes to your local database.

Tip

If you don't see the word "Apply," after clicking Update Local CV, then it means your database is already up-to-date.

---

8. Repeat this process to update the remaining CVs in the list.
9. Once the CVs have been updated, close the ODM Tools.

Now that you've had a hands-on introduction to ODM, let's get to know the raw data that you'll be loading into it.

Getting To Know Your Data

Let's now take a look at the raw data for the workshop, available on your computer at **Workshop\RawData**. These data files were created from real data extracted out of the Lake Champlain Long-term Water Quality and Biological Monitoring Project. Note that the data have been slightly reformatted to facilitate use in this workshop, but much of the terminology used by the data is untouched, giving you a sense of what it takes to transform real data for loading into HIS.

> **Purpose of the Data**
> Some aspects of the data have been modified for use in the workshop. These data are provided solely for use in the HIS workshop, and are not intended to be used for real analysis or decision making.

You should see three files in the RawData folder:
- **sampling_sites.xls** – Excel spreadsheet with locations for all sampling sites
- **LCM_Data.xls** – Excel spreadsheet with the time series of water quality for all sites and variables
- **metadata.txt** – Text file with metadata describing your data

Sites
Open **sampling_sites.xls**. Each row in this file represents a single sampling site and includes the following information:
- **StationID** – Unique internal identifier for a site
- **StationName** – Name of the site
- **Latitude** – Latitude of the site in decimal degrees
- **Longitude** – Longitude of the site in decimal degrees
- **County** – The county that the site is in
- **State** – The state that the site is in

Time Series
Open **LCM_Data.xls**. Each row in this file represents a single water quality measurement at a particular site at a particular point in time. By looking at the dates and times of the measurements, it appears that the data were taken sporadically through time. The spreadsheet includes the following information:
- **Station** – Unique internal identifier for a site
- **Date** – Date that a measurement occurred
- **Time** –Time that a measurement occurred
- **Depth** – Depth at which the measurement was taken
- **Test** – Code for the variable being measured
- **Result** – Value of the variable at the given date and time
- **Method** – Method used to determine the value

Metadata
Open **metadata.txt**. If this data provider wanted to share the data, the information in the sites and time series files alone may not be enough to fully describe the data. Therefore, metadata such as the content of metadata.txt are often provided alongside the actual data files. In this file, you can find information about the nature of the study, the variables involved, and the data source. You'll use this information to help load the data into HIS.

Transformed Output

From these files describing sites, time series, and metadata, you will create transformed files that are formatted for loading into an ODM database using the ODM Data Loader. The transformed files that you will create are:

- Sites
- Sources
- Variables
- DataValues

These are basically the minimum pieces of information you need to describe your data in ODM. Of course, there are plenty of additional types of information you can load into an ODM database to more fully describe your data, but for the purposes of this workshop, you'll just load the above items.

Before performing the transformation, it's imperative that you familiarize yourself with the structure of the Observations Data Model (http://his.cuahsi.org/odmdatabases.html), and the requirements for input files to the ODM Data Loader (http://his.cuahsi.org/odmdataloader.html). Review the information in "Review of CUAHSI-HIS Data Publication Tools," the workshop presentation materials, and the online materials linked above for more information. The following sections describe the transformation procedure.

Creating a Sites File

Information about your sampling sites is contained in the **RawData\sampling_sites.xls** file. You'll transform this to a **TransformedData\sites.csv** file. The fields in the transformed file include:

- SiteCode
- SiteName
- Latitude
- Longitude
- County
- SiteState
- LatLongDatumSRSName

These are some of the field names the ODM Data Loader expects to see when loading information about sites. The first six fields match very well with the data from sampling_sites.xls. Note that you will be adding one field that isn't in sampling_sites.xls: LatLongDatumSRSName. The ODM requires the datum to be stored with the latitude and longitude coordinates of a site. Luckily, you can find that data in the meatadata.txt file. The metadata indicates that the datum used is WGS84. There is already a record for this datum in the SpatialReferences table of ODM (Figure 0-4). The record has a SpatialReferenceID of 3, and an SRSName value of WGS84.

SpatialReferenceID	SRSID	SRSName
0	0	Unknown
1	4267	NAD27
2	4269	NAD83
3	4326	WGS84

Figure 0-4 The WGS84 datum is among the list of coordinate systems in the ODM SpatialReferences table

Being a relational database, the ODM Sites table is expecting a numerical datum ID (in this case, the number 3) to accompany each site record. However, it's easier for us humans to interpret text rather than numbers during the data translation process, which is why the ODM Data Loader allows you to use "WGS84" instead of the number 3 to refer to your datum. The ODM Data Loader will make sure the LatLongDatumSRSName refers to an SRSName in ODM's SpatialReferences table before finalizing the data loading operation. This is one of the advantages of using the ODM Data Loader for loading data – it performs some quality control and maintains integrity of relationships between the tables of ODM during loading.

You'll be following this procedure of referencing information in your ODM database a lot as you figure out how to translate your data to ODM. A quick summary of the procedure is:
1. You need to know how to describe some aspect of your data, such as the datum.
2. You check the ODM database to see if a table already exists to describe that item.
3. You find a match for your item in the ODM table, and use the matching term from ODM as you build your translated data files for eventual loading into the ODM database.

If you don't find a matching item in ODM, you can add it. If you think the item should have been in ODM's CVs in the first place, then you can petition to have the item added to the Master Controlled Vocabularies at http://his.cuahsi.org/mastercvreg/cv11.aspx, although updating those CVs is beyond the scope of this workshop.

Details of how to map from the raw file to the transformed file are below.

Table 0-1 Mapping Raw Data to Sites

Transformed Field	Raw Data Field (from sampling_sites.xls)
SiteCode	StationID
SiteName	StationName
Latitude	Latitude
Longitude	Longitude
County	County
SiteState	State
LatLongDatumSRSName	*Use "WGS84" from ODM SpatialReferences table*

Tip
The ODM Data Loader ignores case in the field names, so Longitude and LONGITUDE are both valid field names.

Tip
If you have any trouble creating the transformed files, you might find it helpful to refer to the solution files in **Workshop\SolutionFiles\TransformedData**.

To create the transformed sites file:
1. Open **Workshop\RawData\sampling_sites.xls**.
2. Save the file in the **Workshop\TransformedData** folder as **sites.csv**. Be sure to select **CSV (Comma delimited) (*.csv)** from the **Save as type** drop down box as you save the file.

> **ODM Data Loader Best Practice – Use CSV Files**
>
> The ODM Data Loader can work with both comma delimited (.csv) files and Microsoft Excel 2003 (.xls) files. However, the author has found that sometimes Excel cell formatting can cause an incorrect interpretation of the data. Therefore, the author recommends saving the transformed files as comma delimited text files, which contain no instructions about how data should be formatted.

3. After saving, click **Yes** if prompted to keep the workbook in CSV format.

Transforming the sites file will be very easy. You'll start by renaming some fields, and then add one new field. Note that you must not misspell any of the field names, or else the ODM Data Loader will not recognize the field.

4. Rename the following fields:
   a. StationID to SiteCode
   b. StationName to SiteName
   c. State to SiteState
5. Add a field called "LatLongDatumSRSName" (without quotes) and calculate all values to be "WGS84" (without quotes).
6. **Save** the file. If prompted about keeping the workbook in CSV format, click **Yes**.
7. **Close** the file. If prompted about saving changes to the file, click **No**. You just saved them, so you should be fine.

Great job! Creating the sites file was a snap since the raw data already had a sites file to begin with. In addition to sites, ODM also keeps a table of data sources. Think of a data source as a group, agency, or institution that operates monitoring sites. Next you'll create a sources file that defines the data source behind your data.

Creating a Sources File

Transforming the sites file was pretty easy since the raw data included a sites spreadsheet that you could start from. However, a file for data sources is not present among the raw data files. Instead, you'll find the information you need to describe the data source in the **RawData\metadata.txt** file. This file is largely extracted from http://www.anr.state.vt.us/dec/waterq/lakes/htm/lp_longterm.htm, albeit condensed a bit for workshop brevity.

The fields in the sources file that you will create include:

- Organization
- SourceDescription
- SourceLink
- ContactName
- Phone
- Email
- Address
- City
- SourceState
- ZipCode
- Citation
- TopicCategory
- Title
- Abstract
- ProfileVersion
- MetadataLink

The information in the transformed file will actually be used to insert data into two tables: Sources and ISOMetadata. The ODM Data Loader will use the last five columns to create an entry in the ISOMetadata table, and relate the information back to your new entry in the Sources table. It knows the appropriate place to put your data in the database.

Details of how to map from the raw file to the transformed file are below.

Table 0-2 Mapping Raw Data to Sources

Transformed Field	Metadata Location	Value
Organization	Source Details	VT Department of Environmental Conservation
SourceDescription	Source Details	The Vermont Department of Environmental Conservation's mission is to preserve, enhance, restore and conserve Vermont's natural resources and protect human health for the benefit of this and future generations.
SourceLink	Source Details	http://www.anr.state.vt.us/dec/dec.htm
ContactName	Source Details	Jill Data
Phone	Source Details	555-555-5555
Email	Source Details	jill.data@champlain.com
Address	Source Details	123 Main Street
City	Source Details	Datatown
SourceState	Source Details	VT
ZipCode	Source Details	15671
Citation	Source Details	P. Stangel, A. Shambaugh, F. Dunlap, (1992-2008), "Sixteen years of water-quality data collection on Lake Champlain, Vermont and New York, United States"
TopicCategory	*Read from ODM TopicCategoryCV*	inlandWaters
Title	Header	Lake Champlain Long-Term Monitoring Program data
Abstract	STUDY DESCRIPTION	The Long-Term Water Quality and Biological Monitoring Project for Lake Champlain has been in operation since 1992. The project is conducted by the Vermont Department of Environmental Conservation (DEC) and the New York State Department of Environmental Conservation with funding provided by the Lake Champlain Basin Program and the two states. The monitoring network includes 15 lake stations representing major lake segments with distinct physical and water quality characteristics.
ProfileVersion	*n/a*	Unknown
MetadataLink	Header	http://www.anr.state.vt.us/dec/waterq/lakes/htm/lp_longterm.htm

To create the sources file:

1. **Using Excel**, create a new file in **Workshop\TransformedData** named **sources.csv**. Be sure to save the file as a comma delimited file. Click **OK** if prompted about saving only the active sheet and **Yes** if prompted about keeping the worksheet in this format.

2. In row 1, type the following column names:
   a. Organization
   b. SourceDescription
   c. SourceLink
   d. ContactName
   e. Phone

A-17

       f.   Email

       g.   Address

       h.   City

       i.   SourceState

       j.   ZipCode

       k.   Citation

       l.   TopicCategory

       m.   Title

       n.   Abstract

       o.   ProfileVersion

       p.   MetadataLink

3. Locate the pertinent information from the metadata file as described in Table 0-2 above and enter it into row 2 of the worksheet.

Note

If you're feeling lazy, you could just copy the data from the solution file. However, looking up information tucked away in metadata is very typical in the data loading process, so don't deny yourself the enriching experience of at least looking up a few of those fields!

---

4. **Save** and **close** the file.

Wow, can you believe we're already half-way through the data transformation process? Two down, two to go! Not even the strictest controlled vocabulary can contain my excitement! Was that too much? Maybe I should apply some database constraints to myself. Ok, seriously, I'm finished now.

At this point, you might be so fired up about data translation that you're already thinking about how to work with that file of water quality time series, but first let's define your variables so that ODM knows what kind of time series data you have.

Creating a Variables File

Your data represents several water quality time series variables. Some information about these variables can be found in the metadata.txt file. Often you'll have to do a bit of legwork to fill in the rest, in order to fully describe your data in ODM.

As a brief summary, measurements for nitrogen, phosphorus, temperature, total suspended solids, and chlorophyll a were taken sporadically in time. Some metadata about these variables can be found in the Variables section of the metadata text file.

The fields in the variables file that you will create include:
- VariableCode
- VariableName
- Speciation
- VariableUnitsName
- SampleMedium
- ValueType

- IsRegular
- TimeSupport
- TimeUnitsName
- DataType
- GeneralCategory
- NoDataValue

VariableName, Speciation, SampleMedium, ValueType, DataType, and GeneralCategory must all conform to terms in ODM controlled vocabularies. This can actually make your life easier because all you have to do is pick the CV term that best describes your variable.

The ODM Data Loader will use the abbreviation for the variable units name to match the variable to a unit in the Units table. This is another example of how the ODM Data Loader performs integrity checks on the data during the loading process.

Details of how to supply values in the transformed file are shown in the table below. The example values are for the nitrogen variable.

Table 0-3 Mapping Raw Data to Variables (for Nitrogen)

Transformed Field	Value	Notes
VariableCode	TN	The term used for nitrogen, from metadata.txt
VariableName	Nitrogen, total	From VariableNameCV
Speciation	N	From SpeciationCV
VariableUnitsName	micrograms per liter	From UnitsName field of Units table, matches units from metadata.txt
SampleMedium	Surface Water	From SampleMediumCV
ValueType	Sample	From ValueTypeCV
IsRegular	FALSE	These are instantaneous measurements made irregularly through time
TimeSupport	0	Use "0" when instantaneous measurements are recorded
TimeUnitsName	Day	The value doesn't really matter for instantaneous data, as long as it matches text in the UnitsName field of the Units table
DataType	Sporadic	From DataTypeCV
GeneralCategory	Water Quality	From the GeneralCategoryCV
NoDataValue	-9999	From metadata.txt

To create the variables file:
1. In the interest of time, a file named **variables.csv** has already been created for you in the **Workshop\TransformedData** folder. This file has records for four out of five variables. You'll add the fifth variable, nitrogen. Open the file with Excel.
2. In row 6, fill in the values for the nitrogen variable. For reference, see Table 0-3.
3. **Save** and **close** the file.

Nice job! Three out of four input files for the ODM Data Loader have now been created. Before you work on the last file, let's go ahead and load the other data into the database.

A-19

Loading Sites, Sources, and Variables into ODM

Before a single time series value is loaded into an ODM database, you must have already loaded information that describes the time series, e.g., the sites, sources, and variables data that you just finished preparing. As you load time series values, the ODM Data Loader will look for the other information related to the time series to make sure it is being appropriately described. If something is missing, the ODM Data Loader will pop up a friendly message basically saying something to the effect of, "Hi there. I see you're loading data, but you haven't told me what the data represent or how to describe it in the database." This is one way the data loader helps to ensure the integrity of your data.

So without further ado, let's run the ODM Data Loader on those transformed files that you've been working on.

1. Open the **ODM Data Loader**.
2. Enter the connection information using the same information as you did for the ODM Tools.
3. Click **Save Changes**.
4. Dismiss the message indicating that the connection was successful.

With the connection set, you are now ready to open the transformed data files and commit them to the database.

5. Click **Open**.
6. Navigate to and open the **Workshop\TransformedData\sites.csv** file. The ODM Data Loader previews the file (Figure 0-5). As indicated in the bottom left corner, the application has recognized that you are loading sites information. It does this based on the field names in the input file you selected.
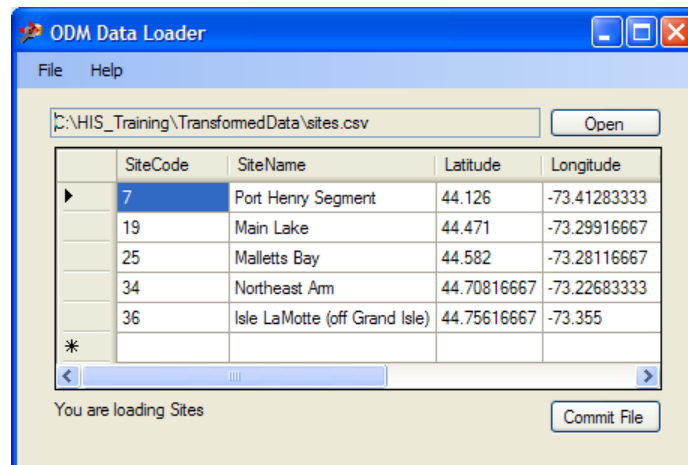
Figure 0-5 Loading sites into ODM

7. Click **Commit File** to write the records to the database. After a moment, the user interface is cleared indicating that the operation has completed.
8. Repeat steps 5-7, this time loading **sources.csv**.
9. Repeat steps 5-7, this time loading **variables.csv**.
10. **Minimize** the data loader when it has finished.

You'll use the data loader again in a moment to load the time series values. For now, take look inside the database to see that data were loaded successfully.

11. Restore **SQL Server Management Studio**. If you closed it, please open and connect to it again.
12. In the **MyWaterData** database, open the **Sites** table.

You should now see your sites in the table (Figure 0-6). If the table is still blank, try closing and reopening the table to refresh it.

SiteCode	SiteName	Latitude	Longitude	LatLongDatumID
19	Main Lake	44.471	-73.29916667	3
25	Malletts Bay	44.582	-73.28116667	3
34	Northeast Arm	44.70816667	-73.22683333	3
36	Isle LaMotte (off...	44.75616667	-73.355	3
7	Port Henry Seg...	44.126	-73.41283333	3

Figure 0-6 Sites information successfully loaded into ODM

Notice the LatLongDatumID field. The ODM Data Loader automatically used the abbreviation for the datum that you provided in the transformed sites file and matched it up with the datum ID from the SpatialReferences table in ODM.

13. Open the **Sources** table. Notice that the data source was automatically assigned a SourceID of 1. Also notice that the Title, Abstract, etc., that you entered earlier has been replaced with a single MetadataID value. If you open the ISOMetadata table, you'll find the additional metadata. The data loader is flexible enough to allow you to load both sources and metadata information from a single file or from two separate files if you had chosen to do so.
14. Open the **Variables** table to see the result of data loading.

With this information in the database, you can now load the time series values.

Creating and Loading a Data Values File

The actual time series of water quality data is stored in raw form at **Workshop\RawData\LCM_Data.xls**. Now that you've loaded metadata about the time series into your ODM database, you're ready to load the time series values themselves.

The fields in the transformed data values file that you will create include:
- SiteCode
- LocalDateTime
- UTCOffset
- OffsetValue
- OffsetUnitsName
- OffsetDescription
- VariableCode
- DataValue
- MethodDescription
- SourceID
- QualityControlLevelID
- CensorCode

A-21

Many of these fields do not have an equivalent in the LCM_Data.xls file. The original data source may not have conceived that their data would be fully described in an ODM database! For some of the additional fields, you will find matching terms or IDs in the ODM database. For example, the ODM Data Loader will match the SiteCode and VariableCode to items in the Sites and Variables tables which have already been loaded. You'll use the SourceID from the Sources table that matches the record you added earlier. Similarly, you can look in the Methods and QualityControlLevels tables to get an idea of what to enter for MethodID and QualityControlLevelID.

Details of how to supply values in the transformed file are shown below.
Table 0-4 Mapping Raw Data to Data Values

Transformed Field	Raw Data Field (from LCM_Data.xls)	Notes
SiteCode	Station	
LocalDateTime	Date and Time	Concatenate the two fields to form a single LocalDateTime field
UTCOffset	"-5"	From metadata.txt, all values are in Eastern Standard Time, which is five hours behind Coordinated Universal Time (UTC), hence the value of -5 for the UTC offset
OffsetValue	Depth	
OffsetUnitsName	"meter"	From metadata.txt
OffsetDescription	"Depth below water surface"	From metadata.txt
VariableCode	Test	
DataValue	Result	
MethodDescription	Method	
SourceID	e.g., "1"	This should be the value of the SourceID for the source data you entered earlier
QualityControlLevelID	"-9999"	"-9999" indicates an unknown QC level, defined in the QualityControlLevels table of ODM
CensorCode	"nc"	From CensorCodeCV table of ODM. "nc" is the default value meaning "not censored."

Like the raw sites file earlier, you'll start with the raw time series file and modify it accordingly.

To create the data values file:
1. Open **Workshop\RawData\LCM_Data.xls** with Excel.
2. Save the file in the **Workshop\TransformedData** folder as **datavalues.csv**. Be sure to save the file as a comma delimited file.
3. Rename the following fields:
    a. Station to SiteCode
    b. Depth to OffsetValue
    c. Test to VariableCode
    d. Result to DataValue
    e. Method to MethodDescription

4. Add a field called "LocalDateTime" and calculate all values to be the concatenation of the Date and Time fields. For example, in row 2, you would use the formula "=B2 + C2".
5. Add a field called "UTCOffset" and calculate all values to be "-5".
6. Add a field called OffsetUnitsName and calculate all values to be "meter".
7. Add a field called OffsetDescription and calculate all values to be "Depth below water surface".
8. Add a field called "SourceID" and calculate all values to be the SourceID that was generated when you created the source information in the database earlier.
9. Add a field called "QualityControlLevelID" and calculate all values to be "-9999".
10. Add a field called "CensorCode" and calculate all values to be "nc".
11. **Save** and **close** the file.

Tip

Quality control levels provide some confidence as to the amount of quality control performed on a dataset. A quality control level of zero (0) indicates raw data, while a quality control level of one (1) indicates quality controlled data. The level of quality control for individual data values is not available for the workshop dataset, so you'll use a value of -9999 to indicate "unknown". For more information on quality control levels, see the ODM design specifications document at http://his.cuahsi.org/odmdatabases.html.

Tip

It's OK to leave the Date and Time columns in the file. Because those are not field names that the ODM Data Loader recognizes, it will just ignore the columns.

Now all that's left is to load the data values into ODM.

12. Restore the **ODM Data Loader.** If you closed the program, please reopen it and connect to the MyWaterData database.
13. Open the **datavalues.csv** file.
14. **Commit** the data.
15. When the data loader finishes, close the program and view the results in the DataValues table in SQL Server Management Studio.

That's it! You've now finished loading all of the raw data into an ODM database. If you are familiar with SQL, you can now write queries to play with other tools in SQL Server to work with the data. But just in case you are not a SQL pro, HIS has developed software called the ODM Tools, which can be used to query and plot graphs of data in an ODM database.

Using the Streaming Data Loader

While the ODM Data Loader is designed to facilitate one-time loading of archived data on disk, the Streaming Data Loader is designed to run as a scheduled task to load data from sensors operating in the field. These sensors typically are connected to a datalogger that handles the recording of data from sensors. Data from the datalogger are sent via a telemetry system to a resource on your computer network, usually in the form of delimited text files. As new data are recorded, the text files are updated with the latest values added to the end of the text file. An alternate scheme is for the telemetry system to create a new text file for each data download. In either case, each row in the text file represents measurements at a single time stamp, and the data values for each variable are stored in separate columns

within the file. A given file is typically associated with only one monitoring site. For an example of this kind of file, see the **sdl.csv** file in your **Workshop\Streaming** folder. More information on the Streaming Data Loader is at http://his.cuahsi.org/odmsdl.html.

In this portion of the exercise, you will use the Streaming Data Loader to load data for a new site that was added to the study. This site only measures nitrogen and phosphorus, and it has the characteristics in Table 0-5.

<div align="center">

Table 0-5 Streaming Data Loader Example Site

Property	Value
Site Code	50
Site Name	Sunset Point
Latitude	44.71
Longitude	-73.25
Datum	WGS84
County	Grand Isle
State	Vermont
Sensor Method	Composite, unfiltered
Sensor Depth	10 meters
QC Level	Raw data
Variables	TN and TP

</div>

Data for this site have already been prepared for you and are located in your **Workshop\Streaming** folder.

To view the example datalogger files:
1. Browse to and open the **sdl.csv** file in your **Workshop\Streaming** folder.
2. Notice the columns for total nitrogen (TN) and total phosphorus (TP).
3. Close the file.

Next you will tell the Streaming Data Loader how to map from data in your file to the ODM database.

To map a data values file using the Streaming Data Loader:
1. Open the **Streaming Data Loader Configuration Wizard** by clicking Setup ODM SDL on your desktop, or by clicking the configuration wizard item in All Programs under CUAHSI HIS.
2. Click the **Add** button (the plus sign) at the top of the window.
3. In the dialog that opens (Figure 0-7), click the browse button next to the Local File text box.
4. In the dialog that opens, browse to and open the **sdl.csv** file in your **Workshop\Streaming** folder.
5. Leave the **Run Every** option to run once per minute. For demonstration purposes, this will ensure that the loader runs each time you tell it to run.

Note – "Run Every" Frequency vs. Scheduled Task Frequency

The "Run Every" option is like telling the data loader how long to rest between attempts to load data. So if I set the "Run Every" option to 1 hour, and I run the data loader seven times within the first hour, then the data loader will run the first time, but for the remaining six times it will say, "Sorry, I'm not even going to look at your data files because I'm still resting." Likewise, if I take a break for the next seven hours, the data loader won't do

---

anything because I didn't tell it to run. If I tell it to run again on the seventh hour, it will say, "More than an hour has passed since I last ran, so I will look at your data files and load new data if present."

You probably don't want to wait around every hour and then click to run the data loader. Fortunately, Windows can run the data loader for you as a scheduled task. You tell Windows how often to run the data loader and then Windows handles the rest. Obviously, you wouldn't set up the scheduled task frequency to be finer than the "Run Every" option or else the data loader would sometimes tell Windows, "I'm still resting" and it won't do any work. You'll set up a scheduled task for the data loader later in this exercise.

6. Specify the connection string to your database as before.
7. Specify that there are column headers on row 1 and that data begin on row 2.
8. Click **Next**.
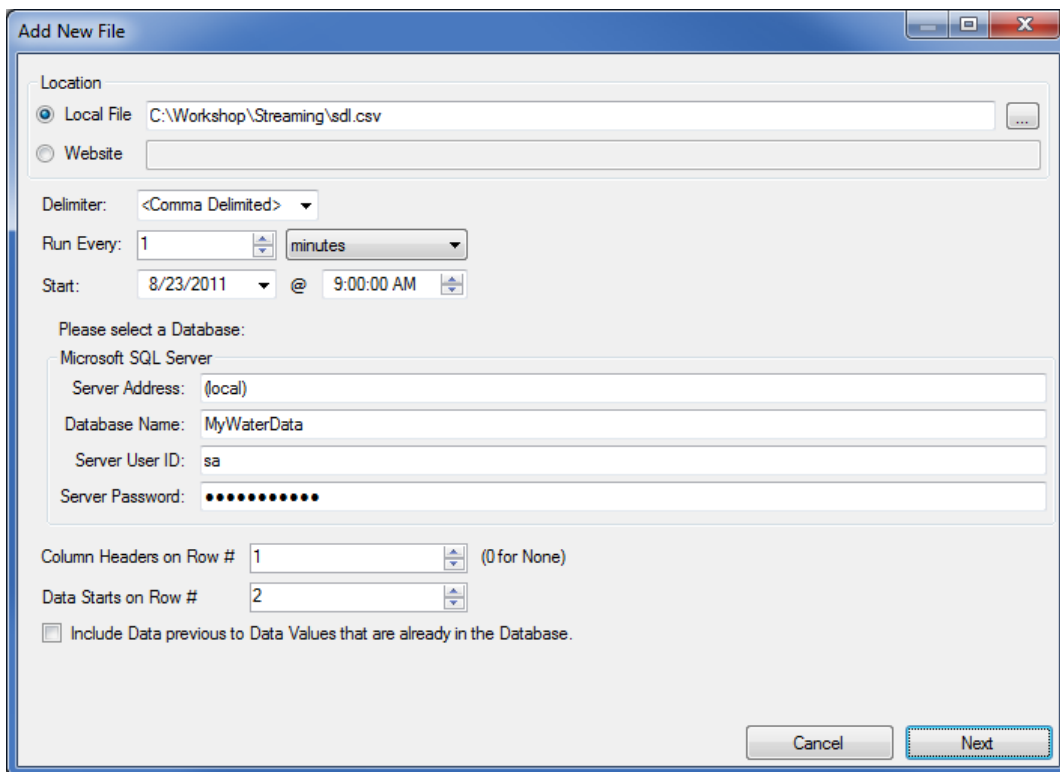
Figure 0-7 Specifying a file and database with the Streaming Data Loader

On the next page of the wizard, you see your data values and some more options to set. You will tell the data loader which column stores your datetimes, and then map your TN and TP columns. In this example, your datetime column stores local datetimes.

9. In the bottom left, set the **Time** option as follows:
   a. Choose the option for **Local Date Time**.
   b. Select the **LocalDateTime** column in the drop down box.
   c. Select **-5** as the time zone. This site is five hours behind UTC time.
   d. Leave the DST box unchecked. The sensor at this site does not use daylight savings time.
10. Near the bottom right, click the **Add** button to map a data value column.
11. Choose **TN** to map the total nitrogen column and click **Next**.

A-25

You've told the data loader that the values in the TN column are time series values. Next you will tell the data loader which site, variable, etc., are associated with values in this column. Notice that the data loader is currently showing you a list of sites that it found in your database. You can choose from this list or create a new site. Since the Sunset Point site is not in the database yet, you will add it.

12. Near the bottom right, click the **Add** button to add a new site to the database.

13. Fill in the site parameters according to values in Table 0-5 (Figure 0-8) and click **OK**. If the information isn't listed in the table, then you can leave it blank.

Figure 0-8 Adding a New Site

The new site now shows up in the list of sites and is selected.

Figure 0-9 Choosing a Site

14. With the Sunset Point site selected, click **Next**.

15. In the list of variables, select **TN** and click **Next**.

A-26

16. Select the "**composite, unfiltered**" method and click **Next**.
17. With the only source record selected, click **Next**.
18. Chose the **Depth below water surface** offset type. Then at the bottom of the window, enter an offset value of **10** (Figure 0-10). Click **Next**.
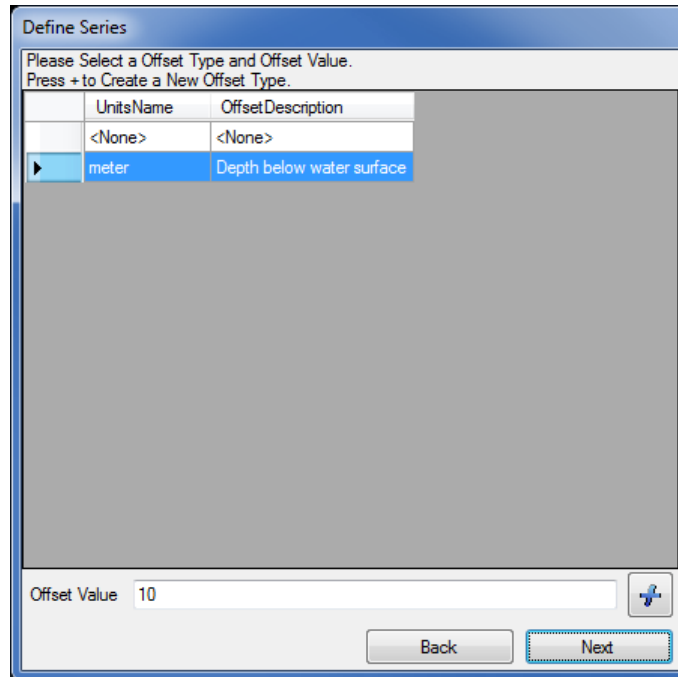
Figure 0-10 Choosing an Offset

19. Select the **Raw data** quality control level and click **Finish**.
20. Repeat the steps above to add the total phosphorus (**TP**) variable.
21. Once TP has been mapped (Figure 0-11), click **Finish** in the Add New File window.
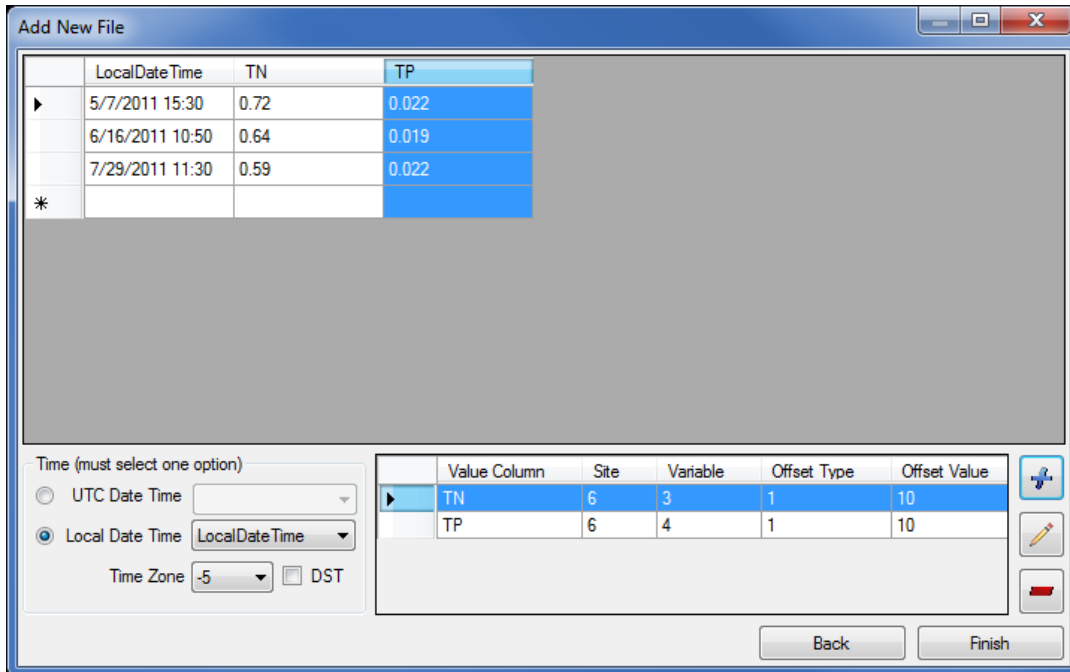
Figure 0-11 Input File Mappings for Streaming Data Loader

The Streaming Data Loader Configuration Wizard now shows a record for the file you just mapped. The configuration options that you specified in setting up this record are stored in an XML file located in your AppData folder, e.g., C:\Users\USERNAME\AppData\Local\CUAHSI\StreamingDataLoader\1.1.2. You can edit this XML manually, which means you could also write a script to automatically create these configuration files. This would be useful if you had a very large network of streaming sensors and didn't want to specify the configuration for each sensor file using the wizard.

You can now execute the data loader by clicking the Run button in the configuration wizard.

To run the Streaming Data Loader from the Configuration Wizard:
1. Click the Run button 🔼 to execute the Streaming Data Loader.
2. After a moment, view the data values in your database using SQL Server Management Studio to see that they have been updated.

Tip

To show just the values for Sunset Point, which was assigned a SiteID of 6 in your database, append the following WHERE clause to the DataValues query in SQL Server Management Studio and click Execute.

```
WHERE SiteID = 6
```

To simulate new values arriving from a sensor, you'll now append the values from sdl-2.csv into sdl.csv. The sdl-2 file contains values for August and September, 2011.
3. Use a program such as Excel or Notepad to copy the values from **sdl-2.csv**, and append those values to **sdl.csv**.
4. Save sdl.csv.

A-28

5. Close sdl.csv and sdl-2.csv.
6. In the Streaming Data Loader Configuration Wizard, click the **Run** button.
7. After a moment, view the data values in your database to see that they have been updated.
8. Close the Streaming Data Loader Configuration Wizard.

The Streaming Data Loader saw that new data values had been added to the sensor file (sdl.csv) and updated the database accordingly. It did not reload the old data values even though they were still present in the sensor file.

Tip

In the case that your monitoring and telemetry system creates a new streaming data file for each datalogger each time data is downloaded, you can connect to multiple local files containing data from the same datalogger by using wildcard characters (i.e., entering 'C:\StreamingData\ThisSite*.dat' will use all files within the C:\StreamingData folder that begin with 'ThisSite' and have a '.dat' extension). All of these files must be formatted exactly the same. The ODM SDL will scan each file each time the update is run for new data to load into the database.

Since the configuration seems to be working correctly with the Streaming Data Loader, you can now set it up to run automatically as a scheduled task in Windows.

Update sdl.csv with even more data:
1. In your **Workshop\Streaming** folder, append the values from sdl-3.csv to sdl.csv.
2. Save sdl.csv.
3. Close sdl.csv and sdl-3.csv.

To set up the Streaming Data Loader to run as a scheduled task:
1. Start **Task Scheduler** (Start | All Programs | Accessories | System Tools | Task Scheduler).
2. Task Scheduler tends to open in a smaller window than it should. **Maximize** Task Scheduler.
3. Click **Action | Create Basic Task**.
4. Enter **Streaming Data Loader** as the name, **Runs ODM Streaming Data Loader** as the Description, and click **Next** (Figure 0-12).
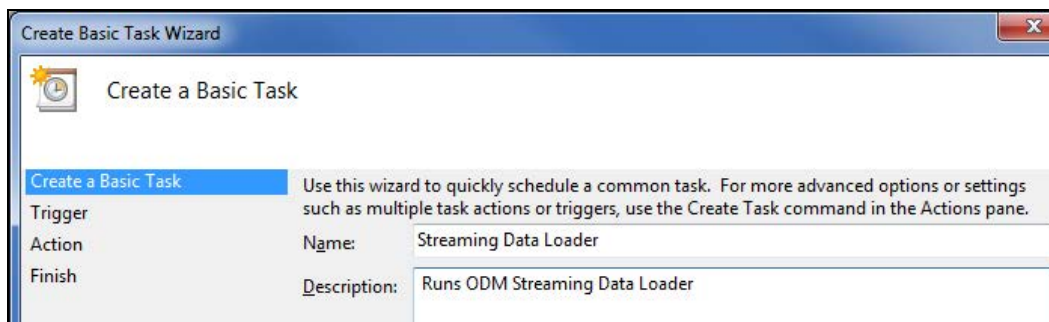
Figure 0-12 Creating a Basic Task with Task Scheduler

5. With **Daily** selected, click **Next**.
6. Choose a time a few minutes from now and click **Next**.
7. With **Start a program** selected, click **Next**.

A-29

8. Browse to the Streaming Data Loader executable (e.g., "C:\Program Files (x86)\CUAHSI HIS\ODM SDL 1.1.2\ODMSDL.exe") and click **Next**.
9. Click **Finish**.

After the time has passed for which the task was scheduled, check your database to see if the new values were appended. Once you've seen that the task completed successfully, you may want to adjust the frequency of the task. Even though the smallest time period available was Daily when creating the task, you can specify a shorter time period in the properties window for a task once it has been created.

To refine the frequency of a scheduled task's execution:
1. In Task Scheduler, click **Refresh** to make sure your task is visible.
2. In the **Active Tasks** section, double-click **Streaming Data Loader** (Figure 0-13).
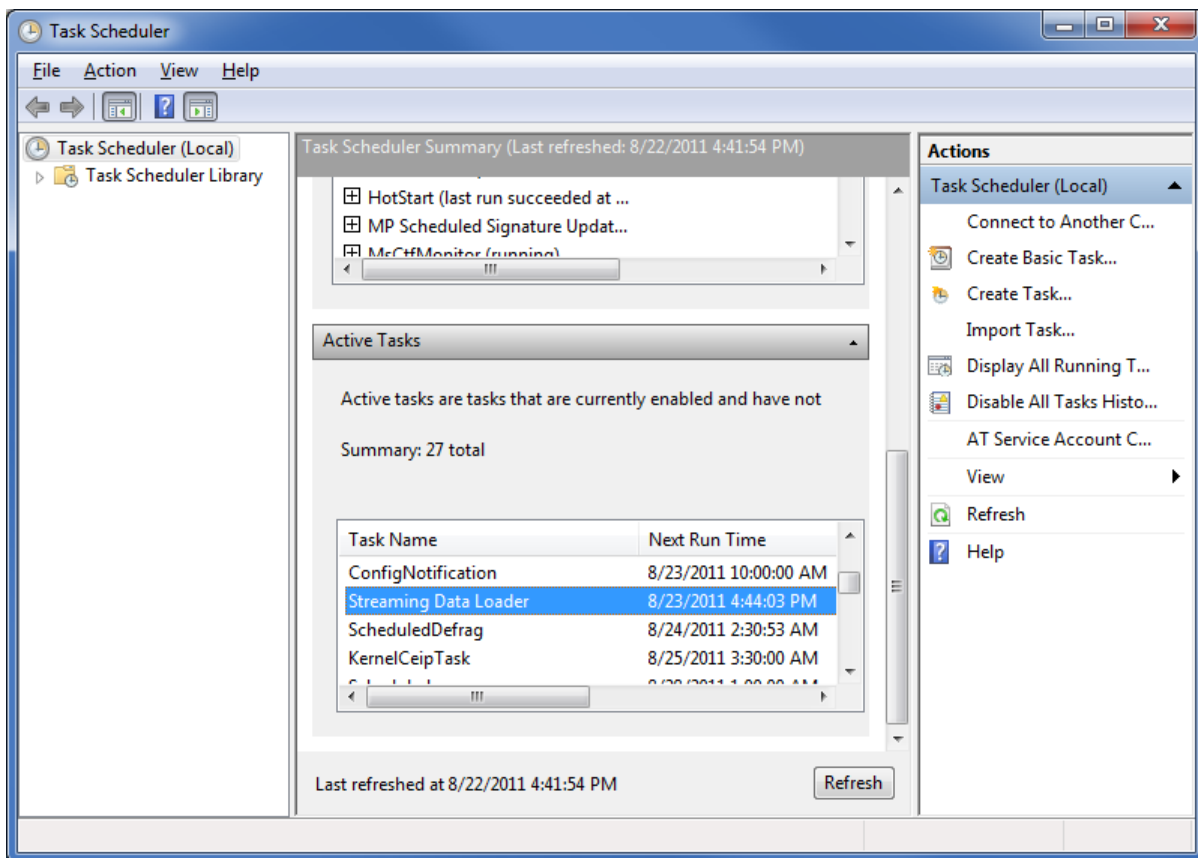
Figure 0-13 Selecting a Scheduled Task

3. Click **Action | Properties**.
4. In the properties window that opens, click the **Triggers** tab.
5. Select the **Daily** trigger and click **Edit**.
6. Place a check next to **Repeat task every**. Leave the default of 1 hour selected.
7. Choose a duration of **Indefinitely** from the drop down list and click **OK** (Figure 0-14).

A-30
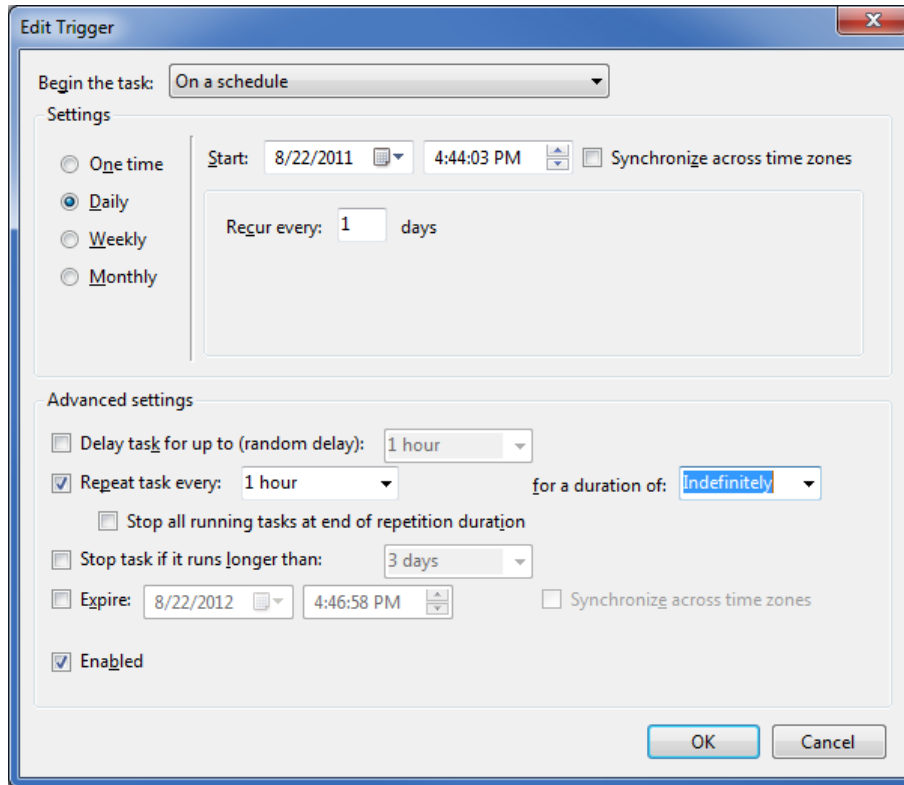
Figure 0-14. Refining the Task Trigger

8. Click **OK** to close the properties window.

Congratulations! You are now a Streaming Data Loader expert! In practice, you would schedule your task to run as frequently as your data files are updated.

With data loaded into the database, you will now view the data using ODM Tools.

For Advanced Participants
ODM Data Loader from the Command Line
To learn more about the ODM Data Loader, read the documentation at http://his.cuahsi.org/odmdataloader.html. Did you know that the ODM Data Loader can be scripted? Open the command prompt and give it a shot.

Work that SQL Magic
Familiar with SQL? Try out some queries to explore your database, or experiment in creating table views. Or, just look around the various tables to get a better sense of what's in ODM. More on ODM can be found in the documentation at http://his.cuahsi.org/odmdatabases.html.

Working with ODM Data

Now that data are loaded into an ODM database, how do we analyze it? Lucky for us, the ODM Tools are specifically designed to query and visualize data within an ODM database. You will use the ODM Tools to examine the contents of the database that you have just created.

To examine your data with the ODM Tools:

1. Open the **ODM Tools**. Notice how the tools automatically connect to the last database you worked with.

The ODM Tools application opens with three tabs visible: Query, Visualize, and Edit. The Query tab is selected by default. On this tab, you specify various filters to search for time series in the ODM database. Take a moment to review the query options.

Our data are fairly uniform in nature, i.e., they all have the same data source, data type, sample medium, etc. This limits the kinds of interesting queries we can perform with our data. But we can still query by site, variable, number of observations, or time period.

Let's query for temperature data as described below.

2. Click the check box to **Query by Variable** (Figure 0-1). This enables you to choose a variable from the variable list.
3. **Select Temperature** in the variables list by left-clicking on it (Figure 0-1).
4. Click the **Query** button in the bottom right corner (Figure 0-1).

The results of the query are shown at the bottom of the application window. You should see several items there, where each one represents a time series of temperature at a particular site.

Tip
You can resize the ODM Tools window to show more query results at the bottom.

---

5. Find the time series of temperature for Main Lake (it's probably the first one in the list). **Right-click the data series** (Figure 0-1). In the context menu, notice that you have options for plotting graphs, editing the data, viewing and exporting metadata, and exporting the data series itself as comma or tab delimited text.
6. In the context menu, click to **View MetaData** (Figure 0-1).
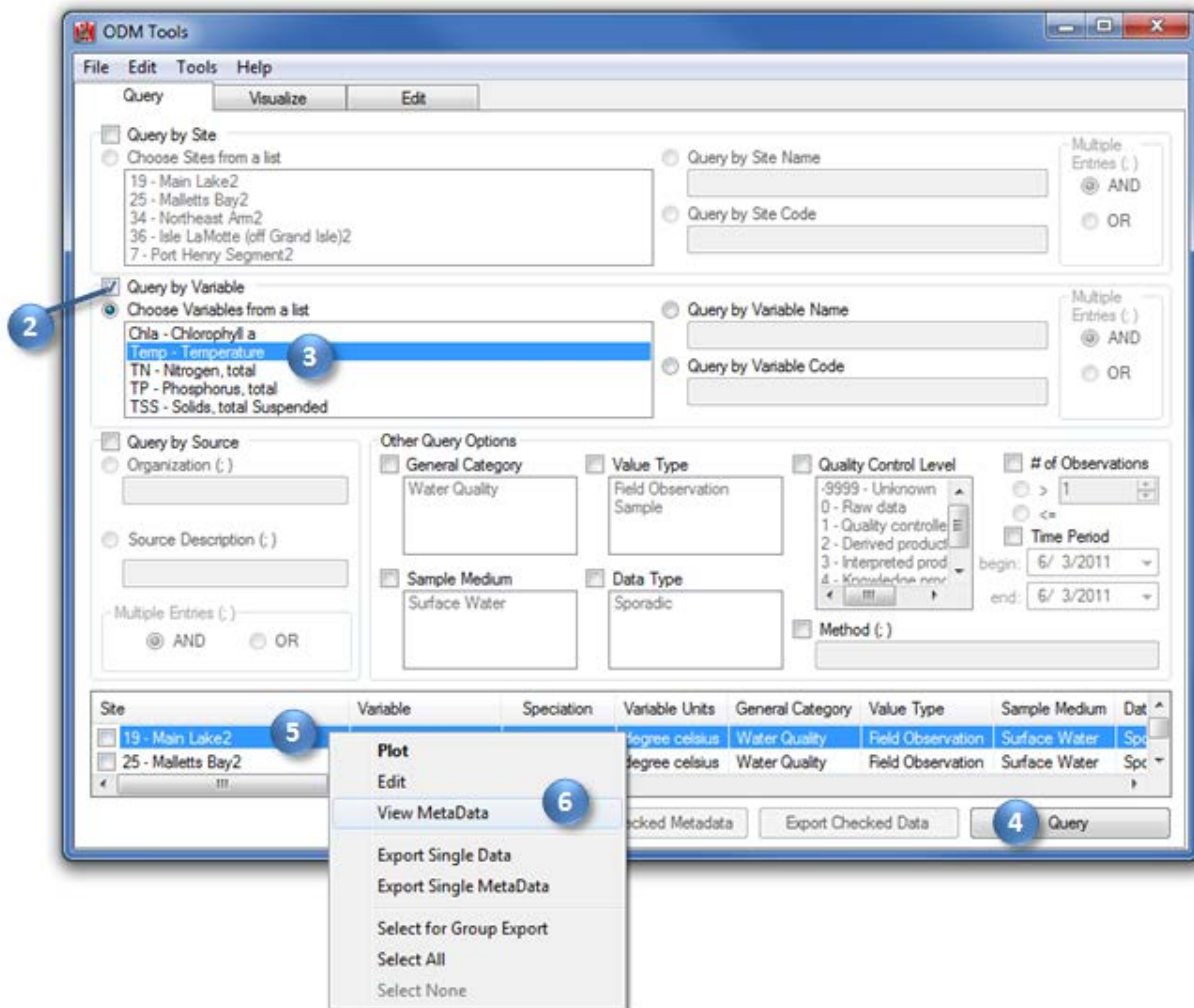
Figure 0-1 Querying data with the ODM Tools

Metadata for the time series are extracted from the database and transformed to XML (Figure 0-2)


```
<VariableInformation>
    <VariableCode>Temp</VariableCode>
    <VariableName>Temperature</VariableName>
    <Speciation>Not Applicable</Speciation>
    <VariableUnits>
        <UnitsName>degree celsius</UnitsName>
        <UnitsType>Temperature</UnitsType>
        <UnitsAbbreviation>degC</UnitsAbbreviation>
    </VariableUnits>
    <SampleMedium>Surface Water</SampleMedium>
    <ValueType>Field Observation</ValueType>
```
Figure 0-2 Metadata exported from ODM using the ODM Tools

7. Right-click the data series again, and this time click to **Export Single Data**.
8. **Save** the file to disk at a location of your choosing.

A-33

9. When the data export is complete, dismiss the message box and open the data file. There are your data again, this time with internal identifiers and other pieces that the ODM Data Loader filled in for you when loading data.

Tip

You can select additional items to be included in the data export by clicking **Tools | Options**.

---

10. **Close** the metadata file and exported data file when you are finished looking at them.

Now let's plot a graph of the data.

11. Right-click the Main Lake data series, and click **Plot**.

You are brought to the Visualize tab, where the ODM Tools plot a graph of the data. Information about the data series you are plotting is shown at the bottom of the application window. The plot has a very "spiky" nature to it. Let's take a closer look to see what's going on.

12. Using the left mouse button, click and hold the mouse button down to draw a box on the plot around one of the "spikes" in the plot to zoom in to that section (Figure 0-3).
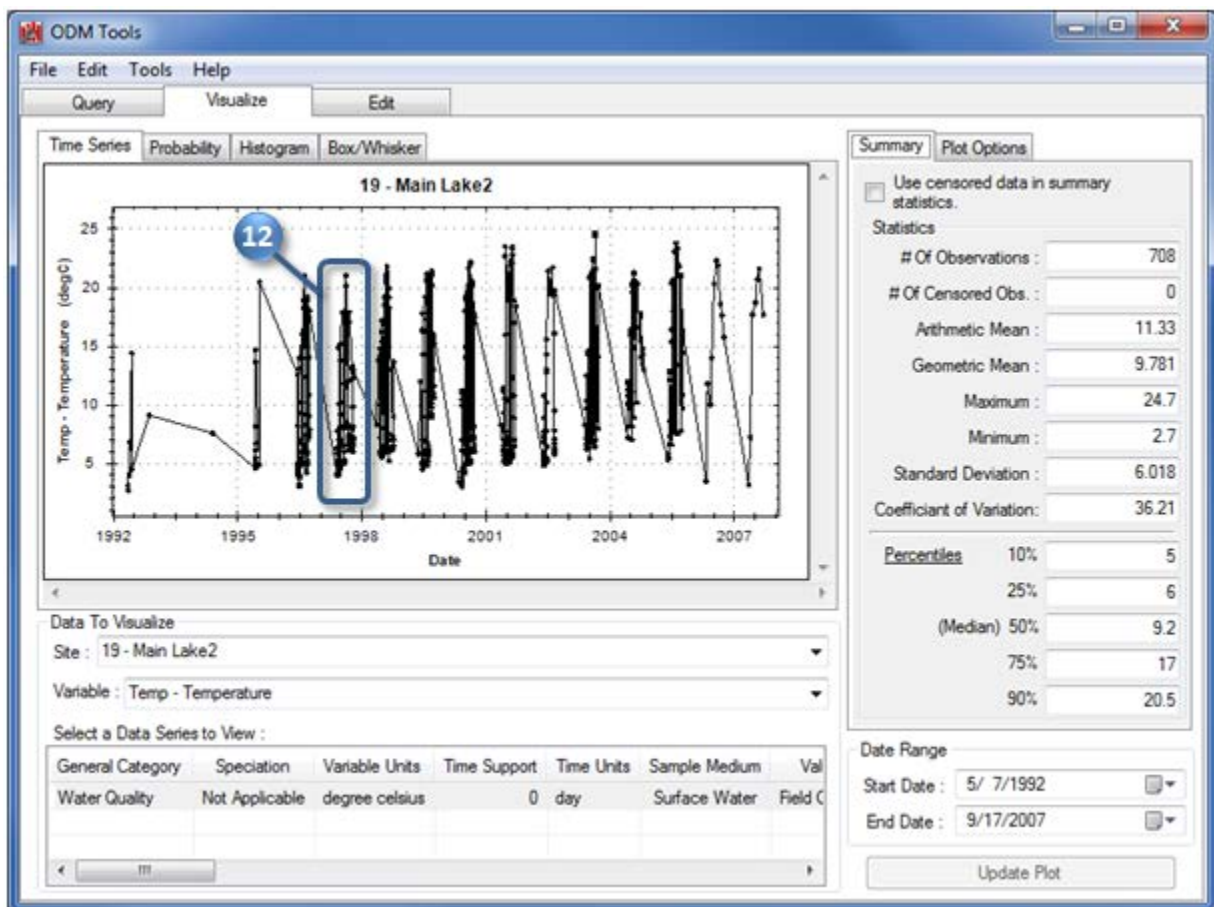
Figure 0-3 Left-click to zoom in on a plot in ODM Tools

Notice that several data points appear to occur at the same point in time (Figure 0-4). This is because several measurements are taken at the same time, but at different depths within the water body.

A-34
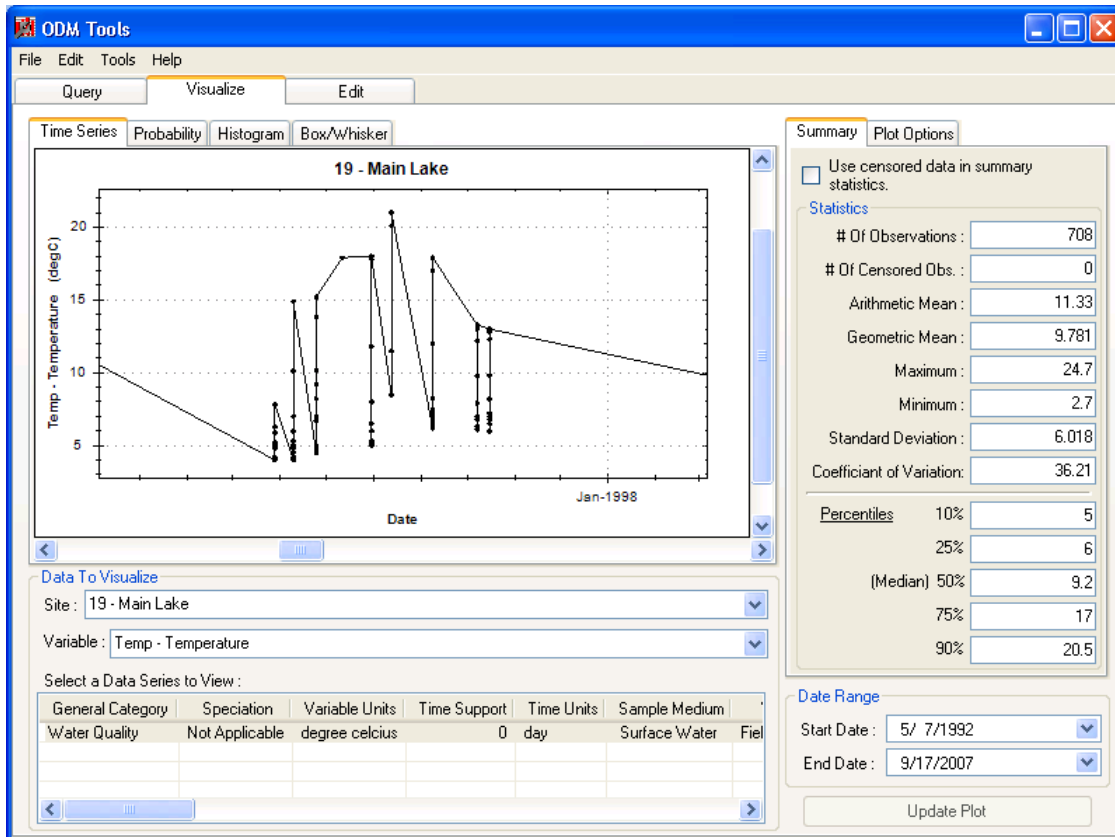
Figure 0-4 Plot of temperature data measured at different depths below the water surface

13. Experiment with the other charting capabilities. You can change plot options, view summary statistics, show a probability plot, show a histogram, and show a box/whisker plot. You can also interact with the plot by clicking on it. Left-click and drag a box to zoom in, and right-click to zoom out or copy and print the chart.

Finally, let's briefly take a look at the Edit tab. You won't be doing any editing for this workshop, but you can at least get a sense of what you can do with the ODM Tools.

14. Click the **Query tab**. Your previous results are still visible at the bottom.
15. **Right-click** the Temperature time series at the bottom, and click **Edit**.

This brings you to the Edit tab and pulls up the time series you selected. Look at the options on the right. You can change individual values or apply a data filter to perhaps look for outliers. Let's try that.

16. Click the option to set a **Value Change Threshold** (Figure 0-5). This option is useful for locating values that differ greatly from the other values around it, which could indicate a sensor malfunction, human error in observation, or some other anomaly.
17. Type a value for the **Value Change Threshold** that will capture some of your data, just to see what happens. For my data, I used a value of 10 (Figure 0-5). Press **ENTER** after you have typed in your value to confirm it. The Apply Filter button should now be enabled.
18. Click **Apply Filter**. Any values that match the filter will be highlighted in red in the plot (Figure 0-5). This shows us values that we may want to check for quality assurance.

A-35
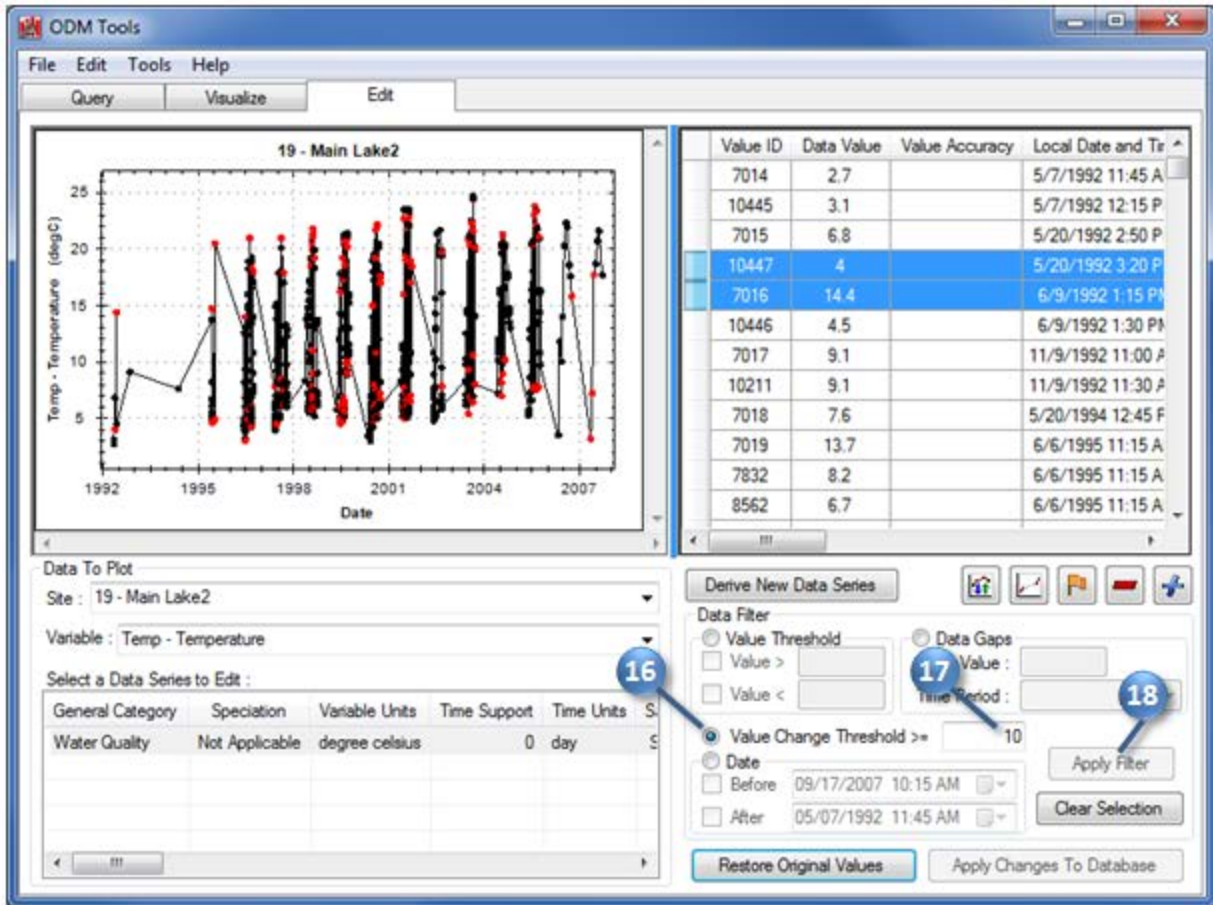
Figure 0-5 Applying a data filter with the ODM Tools

Lastly, let's take a look at options for deriving new data series.
19. Click **Derive New Data Series**.

In the window that opens, you'll see options for applying an algebraic equation, using a daily aggregate function, creating attributes to describe the output, and more (Figure 0-6). These functions are useful for turning your raw information into knowledge products, like when gage height data for rivers are converted to streamflow.
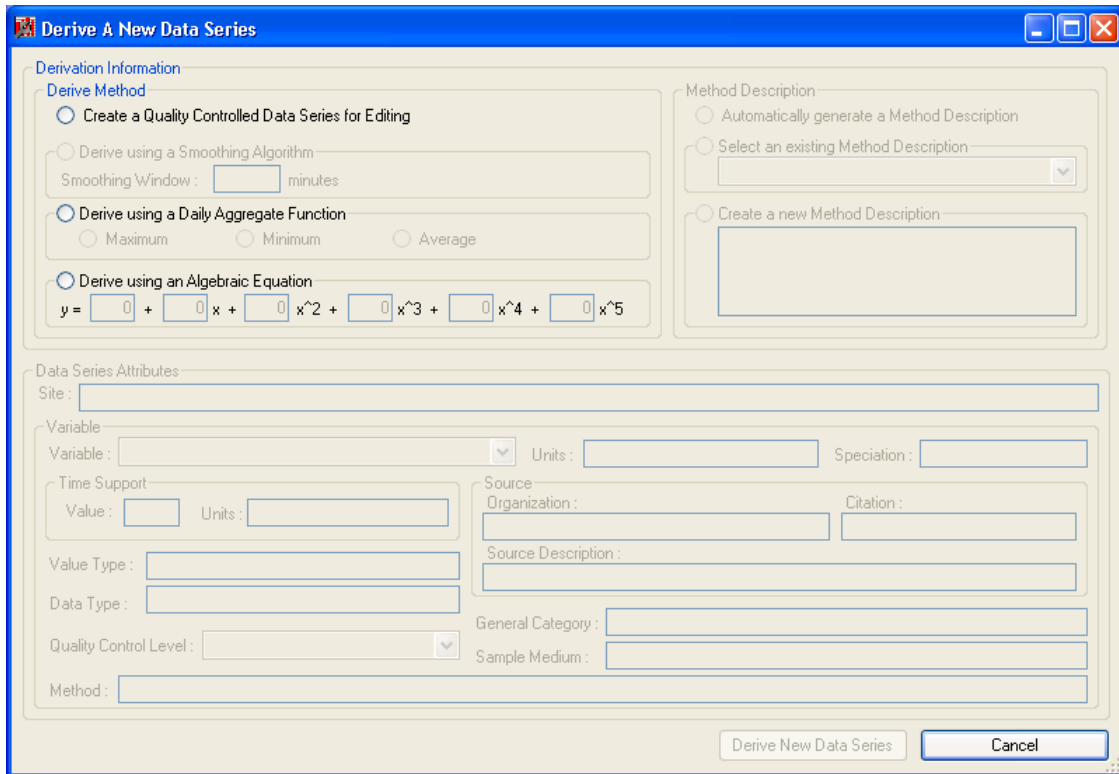
Figure 0-6 Window to derive a new data series in ODM Tools

In your case, because the quality control level is unknown, you'd first have to create a quality controlled data series for editing before the other options would be enabled. However, editing data series is beyond the scope of this introductory workshop, so we'll leave the ODM Tools for now.

20. Click **Cancel** to close the Derive A New Data Series window.

21. **Close** the ODM Tools.

Congratulations! You are now an expert in loading, querying, and visualizing data in an ODM database. For some scientists, this may be all that is desired: a means of storing and working with hydrologic observations data. However, there is often merit in sharing data with a larger community. That's where the WaterOneFlow Web services and HIS Central come in.

Now that you have prepared an ODM database, you will publish the data with a WaterOneFlow Web service to make it accessible online.

For Advanced Participants

Continue playing around with ODM Tools. Try some new queries and deeply explore the options for plotting graphs.

Publishing an ODM Database with WaterOneFlow

As far as your local setup goes, you are now ready for action with your ODM database. But what if you want to share the data with others online? That's where WaterOneFlow Web services come in. WaterOneFlow defines a standard set of queries and a standard output format for accessing data, regardless of whether the data are accessed internally from an ODM database, some other database, or even through another website. Additionally, WaterOneFlow provides a layer of security over your database which makes it less susceptible to hackers than exposing the database itself with public access.

For those who have their own database format for storing data, they must write their own WaterOneFlow Web service to publish their data in HIS. However, you're in luck! HIS includes a free WaterOneFlow Web service specifically designed to work with an ODM database. This means you don't have to write a single line of programming code for your service. You just have to set it up on your computer and tell it to talk to your ODM database.

In this portion of the workshop, you'll download and install a WaterOneFlow Web service to work with your ODM database. The main steps are:
1. Create a SQL Server account that your Web service will use to access your database.
2. Install the WaterOneFlow Web service on your computer.
3. Configure the service.
4. Check the result.

Creating the Webclient SQL Server Account

The WaterOneFlow Web service uses a SQL Server account to connect to your ODM database. This account should have fewer privileges than the one you used for ODM Tools because WaterOneFlow only needs to read data from the database. It does not perform any data creation, updating, or deletion. It is also the public interface to your data, so restricting what the service can do to your database is generally a good idea. In the steps below, you will set up the SQL Server account for WaterOneFlow to use. For this workshop, the account is named "webclient." However, you are free to name the account whatever you choose in your production environment.

To set up the webclient account in SQL Server:
1. If it is not already open, open and connect to **SQL Server Management Studio**.
2. In the Object Explorer on the left, expand **Security | Logins** (Figure 0-1).
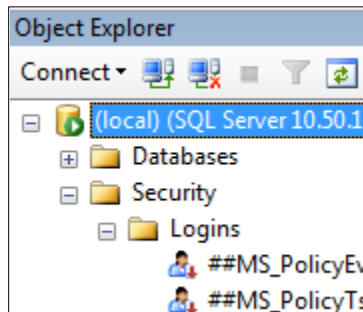
Figure 0-1 Accessing logins for SQL Server

3. Right-click **Logins** and click **New Login**.
4. Type **webclient** as the name (Figure 0-2).
5. Select the **SQL Server authentication** option (Figure 0-2).

6. Type **webclient** as the password (Figure 0-2). Normally you'd want a more secure password, but for the purposes of this workshop, we're keeping things simple and easy to remember.
7. **Uncheck** the option to **Enforce password policy** (Figure 0-2). Again, we're choosing to keep things simple for the workshop. You may want to enforce the SQL Server password policy for your own installation.
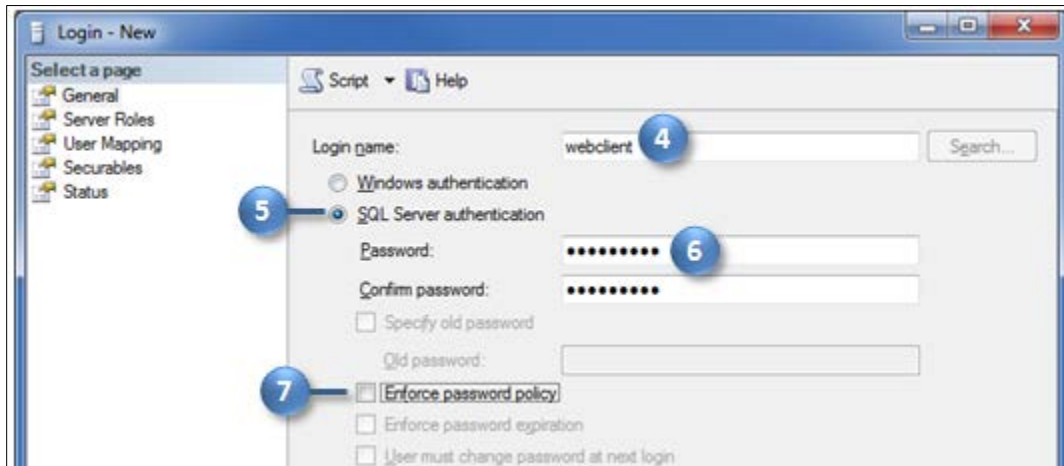
Figure 0-2 Creating a webclient SQL Server login

Note

Suppose you've already published one ODM database with WaterOneFlow, and have now created a second database that you want to publish. Since the webclient account was created when you published the first database, you do need to repeat those steps. You can start from the next step with User Mapping.

With the webclient login created, you will now add the MyWaterData database to the list of databases that the login can access.

8. In the **Select a page** pane on the left, click **User Mapping** (Figure 0-3).
9. Place a check next to **MyWaterData** and click **OK** (Figure 0-3).
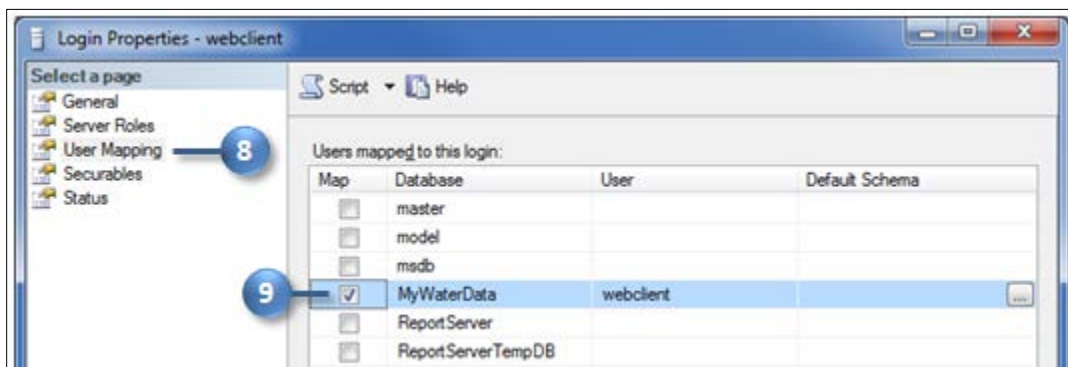
Figure 0-3 Allow the webclient to access the database

You're almost finished. The last thing to do is allow the webclient account to perform Select operations on the database. The Web service needs this in order to properly query the database. You allow this permission on the properties page for the database itself.

10. In the Object Explorer on the left, expand Databases until you see your MyWaterData database.
11. Right-click **MyWaterData** and click **Properties**.

A-39

12. In the Select a page pane on the left, click **Permissions** (Figure 0-4). You should see the webclient account in the list of users for the database.
13. In the list of permissions, scroll to the bottom and place a check in the **Grant** column for **Select** (Figure 0-4).
14. Click **OK** to close the dialog (Figure 0-4). You may also close SQL Server Management Studio.
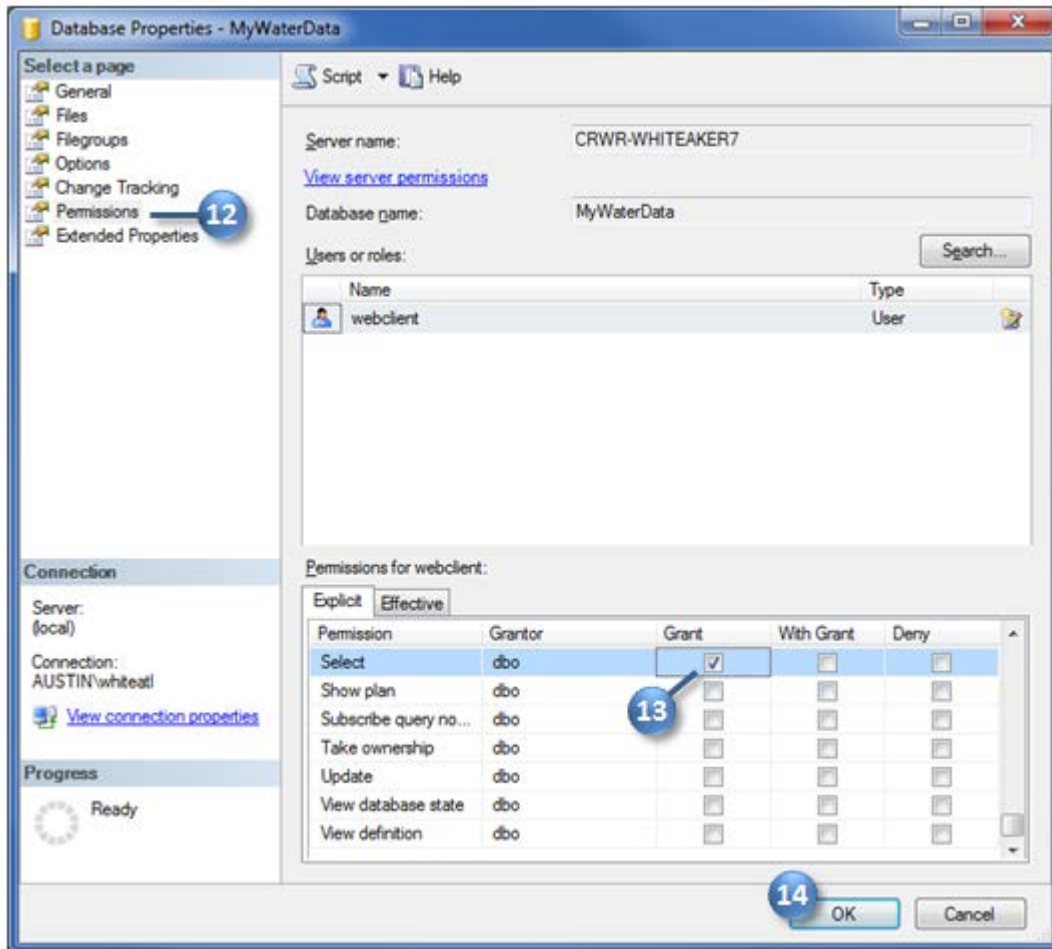
Figure 0-4 Setting permissions for webclient on the database

Great job! The database now has a special login for working with the Web service. The Web service will use this login when connecting to the database. Next, you will install and configure the Web service.

Installing a WaterOneFlow Web Service

To install a WaterOneFlow Web service, you'll download a copy from the HIS website and put it in a folder on your computer. Then you'll use Internet Information Services Manager to configure the service.

To download the WaterOneFlow Web service:
1. In a web browser, navigate to http://his.cuahsi.org.
2. Under **Quick Links** on the right, click **WaterML Web Services**.
3. Click the link to **Download ODM WaterOneFlow Web Services (current version)**.
4. Save and open the zip file.

The zip file contains several files related to WaterOneFlow. The Web service application is in the WebApp folder. You'll copy this folder to the default folder for web applications on your computer.

5. From the zip file, copy the **WebApp** folder to **C:\Inetpub\wwwroot**.

6. Rename the WebApp folder to **MyDataService**. You can name the folder whatever you want, but we'll use MyDataService for this exercise.

Now you'll use Internet Information Services (IIS) Manager to configure. This is a program you can use to manage web applications on your server. To isolate WaterOneFlow from your other web applications, you'll set up an application pool for WaterOneFlow to run in. This pool will use the Classic managed pipeline mode since this is what WaterOneFlow was built to use.

To set up an application pool:

1. Open **Internet Information Services (IIS) Manager**.

2. In the Connections box on the left, expand the tree view until you see **Applications Pools**.

3. Right-click **Applications Pools** and click **Add Application Pool** (Figure 0-5).
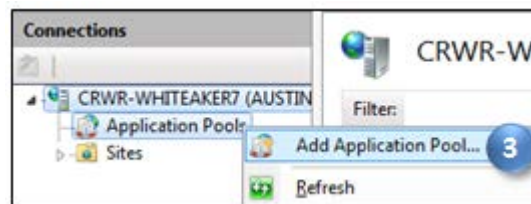
Figure 0-5 Adding an application pool

4. In the Add Application Pool dialog, specify these options (Figure 0-6):

   a. Name: **WaterOneFlow**

   b. .NET Framework version: **.NET Framework v2.0**

   c. Managed pipeline mode: **Classic**
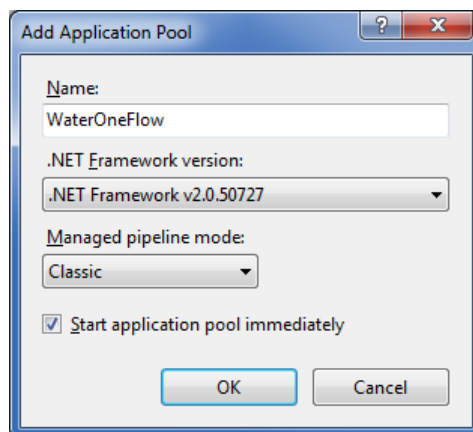
Figure 0-6 WaterOneFlow application pool settings

5. Click **OK** to close the dialog.

Tip

If you are creating multiple instances of the WaterOneFlow web services on your server because you have

multiple ODM databases, you can reuse the WaterOneFlow application pool that you just created for those services.

Now you are ready to install the service and configure it.

To install the WaterOneFlow Web service:
1. In IIS, expand the tree view to **Sites | Default Web Site | MyDataService**. MyDataService is currently just a folder that IIS has found. Next you'll tell IIS that this folder is actually a web application.
2. Right-click **MyDataService** and click **Convert to Application**.
3. In the Add Application dialog, click **Select** (Figure 0-7).
4. In the Select Application Pool dialog, choose **WaterOneFlow** (Figure 0-7).
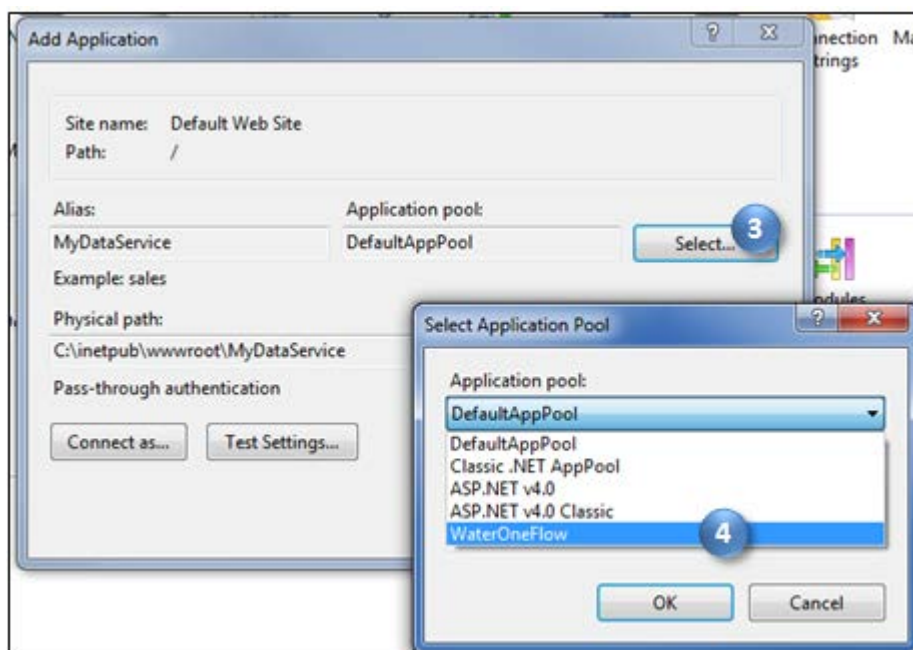
Figure 0-7 Selecting the WaterOneFlow application pool

5. Click **OK** to close the Select Application Pool dialog.
6. Click **OK** to close the Add Application dialog.

With the service installed, you'll next configure the service to tell it how to connect to your ODM database.

To configure the WaterOneFlow Web service:
1. Left-click **MyDataService** to select it (Figure 0-8).
2. Under the ASP.NET icon group in the middle section of the IIS Manager window, double-click **Connection Strings** to open the connection strings editor (Figure 0-8).

A-42
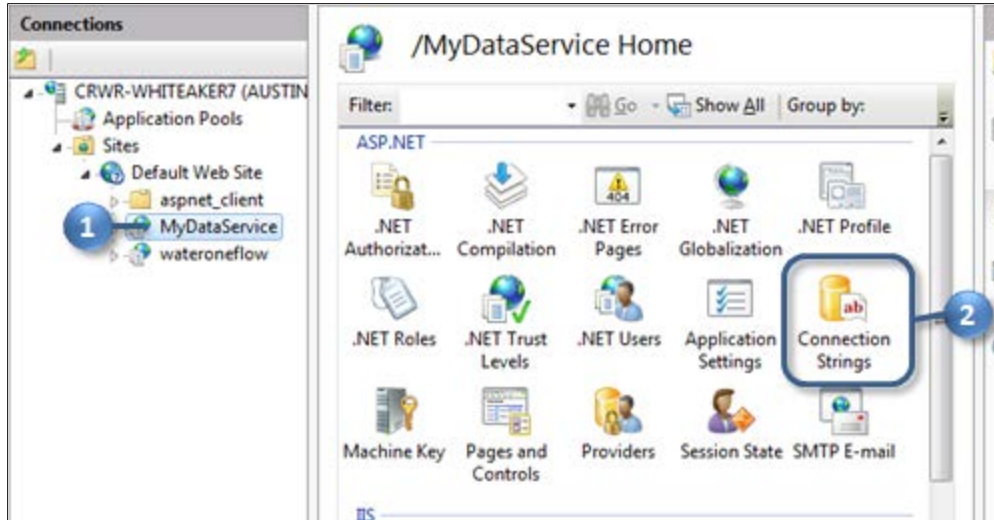
Figure 0-8 Accessing service connection strings

3. Double-click **ODDB** to set the database connection string.
4. Set these parameters in the Edit Connection String dialog:
    a. Choose **SQL Server**.
    b. Server: **(local)**
    c. Database: **MyWaterData**
    d. Credentials: Choose to **Specify credentials**.
5. Click **Set** to set the SQL Server credentials.
6. Set these credentials (Figure 0-9):
    a. User name: **webclient**
    b. Password: [same as the webclient password you established earlier]
    c. Confirm password: [same as above]
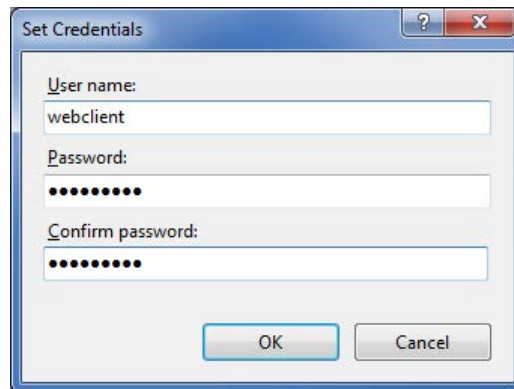
Figure 0-9 SQL Server credentials for WaterOneFlow to use

7. Click **OK** to close the Set Credentials dialog. Your Edit Connection String dialog should now look similar to the one in Figure 0-10.

A-43
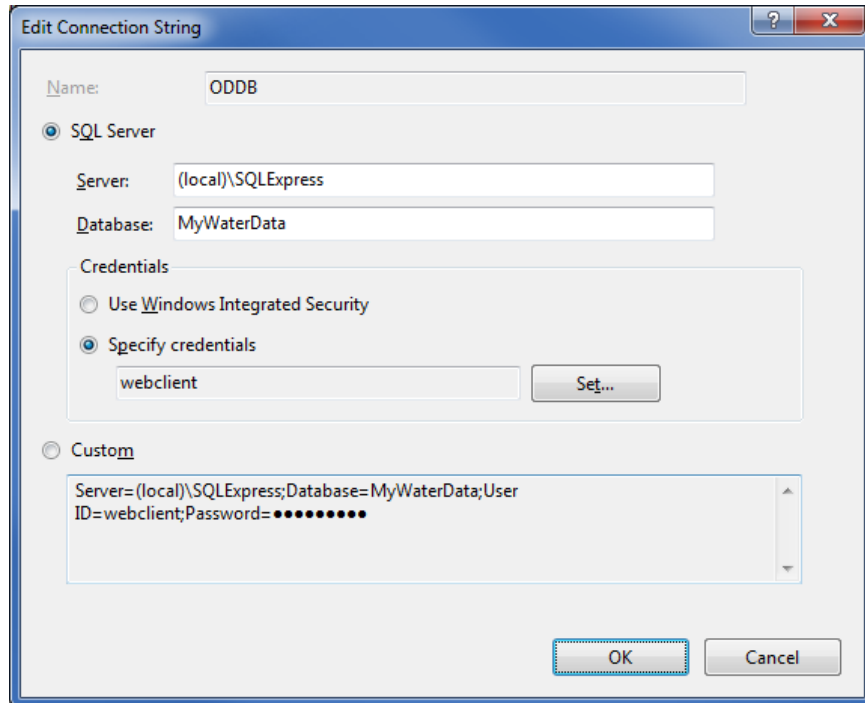
8. Click **OK** to close the Edit Connection String dialog. Now you'll set some other parameters about the service.
9. Left-click **MyDataService** to select it.
10. Under the ASP.NET icon group, double-click **Application Settings**.

There are several items you could set here. In your production environment, you should set at least the following:

-   contactEmail – The email address of the person responsible for administering the service.
-   EncryptConnectionString – You'll probably want to set this to True to encrypt the database connection string.
-   serviceDescriptionPage – URI to a web page providing more information about the service.
-   network – Unique code for the observations network that all monitoring sites in your service belong to. For this exercise, all sites belong to a fictitious "LakeChamplain" monitoring network.
-   vocabulary – Unique code for the vocabulary to which your variables belong to. In this workshop, all variable descriptions come from the Vermont Department of Environment and Conservation, so you'll use the code VTDEC.

Additional settings are described in documentation at http://his.cuahsi.org/wofws.html. For this workshop, you'll just set the network and vocabulary parameters. The other parameters are set in a similar manner.

11. Double-click **network**.
12. In the Edit Application Setting dialog, change the value from ODM to **LakeChamplain**. Please add the number of your computer to the network name to produce, e.g., LakeChamplain27 (Figure

0-11). These numbers will help keep your service distinct from that of other participants in this workshop.
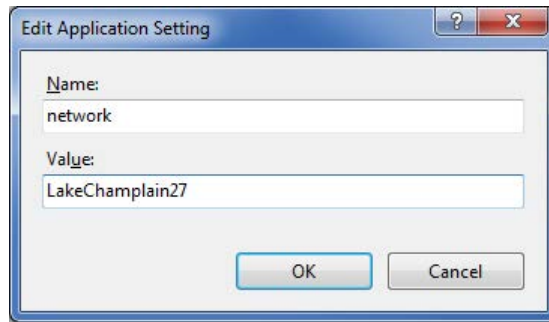
Figure 0-11 Setting the network name

13. Click **OK** to close the Edit Application Setting dialog.
14. Similarly, change the Value of vocabulary to **VTDEC**.

With the service configured, you are now ready to test it.

Testing the Web Service in a Web Browser

You'll test the service by exploring it in a web browser and by using HydroExcel.

To test the service in a web browser:
1. In a web browser, navigate to http://localhost/MyDataService.

Tip
The keyword "localhost" tells your browser to look on your own computer for the web page that you are attempting to access.

The Web service displays this web page when accessed with a browser. Near the top of the page is a link for Database Test Page that you can use to test the service's connection to your database.
2. Click **Database Test Page**.

In a moment, you should see some example data from your database (Figure 0-12). If you see an error message, you can use the message to track down the cause of the error.

This should display up to 10 sites

SiteCode	SiteName
19	Main Lake2
25	Malletts Bay2
34	Northeast Arm2
36	Isle LaMotte (off Grand Isle)2
7	Port Henry Segment2

Figure 0-12 WaterOneFlow database test page

Excellent! The Web service should now be up and running. Let's test the service using HydroExcel.

Testing the Web Service with HydroExcel

At this point in the data publication process, you should be able to give someone the URL to your Web service, and they should then be able to query data from it using any software that communicates with Web services. CUAHSI HIS includes free software called HydroExcel that lets you access WaterOneFlow Web services from within a Microsoft Excel spreadsheet. In this portion of the workshop, you'll use HydroExcel to extract data from your Web service.

To test the Web service with HydroExcel:
1. In a web browser, navigate to http://his.cuahsi.org.
2. Under **Quick Links** on the right, click **HydroExcel**.
3. Click the link to download the **Microsoft Office 2007/2010 version**.
4. For this exercise, save HydroExcel to your Desktop.
5. **Open** the file when it has finished downloading. If prompted, click to enable editing and enable content.

Note

HydroExcel requires the free HydroObjects software to be installed. This software was installed on the workshop computers prior to the workshop. If you are working from a different computer, you can find the installation file at http://his.cuahsi.org/hydroobjects.html.

The worksheets in HydroExcel call methods from a WaterOneFlow Web service to query data and write the result into the spreadsheet. For an in depth tutorial on HydroExcel, see the software manual on the HIS website. For this workshop, we'll just do a quick test to download a list of sites and what variables they have, and also a time series for a given variable at a given site.
6. Activate the **Data Source** worksheet. (Click "Data Source" at the bottom of HydroExcel.)

On the Data Source worksheet, you tell HydroExcel which Web service you want to work with by inputting the URL address of the service next to the box that says "WSDL Location". Some URLs are already listed in the spreadsheet, but you will have to locate your own URL that points to the Web service you just created. Since you're still using the same computer on which the Web service is installed, you could just use the localhost URL. However, let's use the actual IP address of your computer since that's how other people will be connecting to it.
7. Locate your **IP address**. You can find your IP address from the command line:
    a. Click **Start**.
    b. Type **cmd.exe** and press ENTER.
    c. In the window that opens, type **ipconfig** and press **ENTER**.
    d. Write down the numbers that appear to the right of **IP Address**. These numbers look something like "129.116.104.171". Note that this address will be different for each computer.
    e. **Close** the command window.

From this point forward, if you see the text "[YOUR_IP]", please replace it with your IP address. For example, if you are asked to enter the address "http://[YOUR_IP]/MyDataService", and your IP address is 129.116.104.171, please enter "http://129.116.104.171/MyDataService".

---

8. In a web browser, navigate to **http://[YOUR_IP]/MyDataService**.
9. Click the link for the **1.1 version** of the service.
10. Click the link for **Service Description**. This takes you to the WSDL for your Web service.

Tip

The service WSDL (Web Services Description Language) is where your Web service defines what it can do and how programs can interact with it. It is designed for programs to read, so don't worry if you can't make sense of it. When a program accesses your Web service, it will read the WSDL and know exactly how to send requests to it, and will also know what format of output to expect back.

---

11. From the address bar of your web browser, **copy the URL** of the WSDL and paste it into the cell next to the cell that says **WSDL Location** on the Data Source worksheet of HydroExcel.
12. Activate the **Series Catalog** worksheet.
13. Change the option to **Create and open KML file after download** to **TRUE**.
14. Click **Get Series Catalog**.

After a moment, your spreadsheet is updated with information about the sites and variables measured at the sites. Also, Google Earth opens to show the site locations.

15. Browse around Google Earth to see your site locations. Click on placemarks for sites to see information associated with the sites.
16. Switch back to **HydroExcel**.
17. Dismiss the message box indicating that the download is complete.

Take a look at the information in the Series Catalog. You not only get information about the location of sites in your database, but also the variables measured at those sites. Notice that the start date, end date, and number of records of time series observations are included with each variable. You'll use this information to download a time series for one of the sites.

18. Locate a site and variable for which you'd like to download time series data. For my screenshots, I'm going to get **temperature** data at **Main Lake**, so you may want to do the same.
19. In the **Series Catalog** worksheet, **right-click** anywhere on the row for the site and variable that you want.
20. In the context menu that opens, point to **HydroExcel**, and then click to **download the time series**.

You are brought to the Time Series worksheet, where HydroExcel filled in the parameters to make the request for the data, called the Web service, and populated the result in the spreadsheet.

21. Dismiss the message box indicating the download is complete.

Notice there are several measurements taken at a given datetime, but at different depths below the water surface (Figure 0-13).

A-47

Get Values Options	☑ Ignore NoData Value		Variable Name	Temperatur
Site Code/Location	LakeChamplain27:19		Units	degree celc
Variable Code	VTDEC:Temp		SampleMedium	Surface Wa
Start Date	5/7/1992 11:45		Site Name	Main Lake
End Date	9/17/2007 10:15		Latitude	44.471
			Longitude	-73.299166
Get Values			Data Source	http://129.1
			Obtained	5/15/2009 1
			Ignore NoDataValue	TRUE

DateTime	Value	Qualifier	Offset	Offset Unit
5/7/1992 11:45	2.7		2	meter
5/7/1992 12:15	3.1		100	meter
5/20/1992 14:50	6.8		2	meter
5/20/1992 15:20	4		103	meter

Figure 0-13 Time series returned from the Web service

Kudos to you! You've come a long way, and now have successfully made your data available online using WaterOneFlow. Do you feel the magic in this moment? I certainly do. You're almost finished with the entire data publication process. The last major step is to register your service with HIS Central so that others can discover it.

For Advanced Participants

Fun with Pivot Tables

Have you worked with pivot tables in Excel before? They are nifty. Suppose you want to show temperature data 50 feet below the water surface in the lake. On the Statistics and Charts worksheet, you can drag Offset to the Report Filter box, and then choose "50" after clicking on the drop down arrow next to Offset in the Pivot Table Field List. Experiment with other ways of summarizing the data with pivot tables.

By the way, pivot tables and charts are a native part of Excel functionality, and aren't just limited to HydroExcel. You may find pivot tables useful in other spreadsheets that you've created.

Registering Your Service at HIS Central

Although your Web service is now online, there's still the question of how people will find out about it so that they can use it. That's where HIS Central comes in. HIS Central is a special server maintained by the HIS team which keeps a catalog of WaterOneFlow Web services. When you publish a WaterOneFlow Web service, you should register it with HIS Central to make it discoverable. More than just a listing of WaterOneFlow services, HIS Central performs these functions:

- Provides detailed information about your service, including contact information, abstract, and areal extent of your sites.
- Supports translation from your variables to an ontology of common hydrologic concepts. This facilitates easy searching for variables in your service, especially by those who aren't familiar with what your service has to offer.
- Maintains a catalog of sites and variables available in all registered services, enabling fast searching for data from multiple data providers.

The registration process involves these three key steps:

1. Add an entry at HIS Central for your Web service.
2. Tag variables from your service to the hydrologic concept ontology.
3. Check that your sites show up in HydroDesktop.

Once you've completed these steps, the entire publication process will be complete!

Workshop Links for HIS Central and HydroDesktop

Rather than register your service at the official **HIS Central** (http://hiscentral.cuahsi.org), you'll be using a special sandbox version set up just for workshop use at http://water.sdsc.edu/hiscentralsb/. That way, we don't clutter the production system with datasets that we're just using for demonstration purposes. When you're ready to contribute to HIS with your own data and services, then please use the official HIS Central link.

Adding Your WaterOneFlow Web Service to HIS Central

For this portion of the workshop, you will add an entry for your Web service with HIS Central. You will not only give HIS Central the URL to your service WSDL, but also supplementary information about the service.

To register a WaterOneFlow Web service with HIS Central:

1. In a web browser, navigate to the workshop HIS Central at http://water.sdsc.edu/hiscentralsb/.
2. At the top of the page, click the link to **Login**.
3. Log in with your credentials. If you currently in the HIS workshop, your instructor will provide you with credentials. Otherwise, feel free to register your own free HIS Central account.
4. Once you are logged in, near the top of the page, click the link to **Add Data Service**.
5. Add service details (Figure 0-1):
    a. To help distinguish your service from others in the workshop, the service title should be unique among all registered services. Therefore, please enter "Lake Champlain Workshop" (without quotes) and your computer number as the **Service Title**, e.g., Lake Champlain Workshop 27.
    b. Input the **Network Name** that you assigned to your service earlier, e.g., LakeChamplain27.
    c. Input your WSDL address, i.e., http://[YOUR_IP]/MyDataService/cuahsi_1_1.asmx?WSDL, as the **Service WSDL**.
    d. Click the link **I have read and agree to the Data Service Agreement** to see the data service agreement. The link opens in a new window.
    e. Be sure the box is **checked** next to the link for the data service agreement.

Figure 0-1 Registration page for a new service at HIS Central

6. Click **Next**. This brings you to the Data Service Details page.

You now have an entry for your service in the system at HIS Central. However, there are still some steps to take before you make it public. Let's continue editing details of this service.

7. About half-way down the page, click **Edit Details**.

This brings you to a page that lets you edit the description of your service. For the workshop, we'll only be adding a few items, but if you are a fast typist, feel free to add more!

8. Add some info to the details page, such as the following:

a. Organization: HIS Workshop
b. Name: Workshop Participant
c. Citation: Doe, Jane (2011). Water quality monitoring data from the Lake Champlain Long-Term Monitoring Program. VT Department of Environmental Conservation.
d. Abstract: Here's a short abstract.

Note

When you register a service at HIS Central, please use the following citation format:

Name (Year). Type of Data. Network. Collecting Organization.

---

9. Be sure to check the box next to **Is service public**. If you don't check this box, sites from your service will not show up in applications like HydroDesktop.
10. Click **Update**.

This brings you back to the Data Service Details page.

Now let's add images that will be associated with your service. You'll add a logo for your organization that users will see when they view details for your service at HIS Central, and a small icon that represents a site location that will appear in the HydroDesktop map.

11. Change the images for your service.
    a. Click **Change Images**.
    b. For the organization icon, browse to the **Workshop\Images** folder on your computer, and open the **OrganizationIcon.gif** file.
    c. Click **Upload** to upload the image once you have located it. The web page will be updated with your image once the upload is complete.
    d. For the map icon, browse to and add the M**apIcon.jpg** file, located in the same directory as the organization image.
    e. Click Upload to upload the image.
    f. Click **Back**.

At this point, you could add additional contacts, links, and descriptions, which would show up as part of your service's details. This is not required for the workshop. Therefore, your data are now ready to be harvested.

What's harvesting, you ask? Recall that HIS Central keeps a catalog of all of your sites and variables, which enables fast searching across all registered services. HIS Central creates this catalog by calling various methods from your WaterOneFlow Web service. This is called data harvesting. When you request a data harvest, an HIS Central administrator is notified and will trigger a harvest of your data. To keep the system from being bogged down by numerous harvest requests, only an HIS Central administrator can trigger a harvest. Harvesting is essential, because HIS Central must know what variables are available in your service before you can tag those variables to concepts in the hydrologic ontology.

Normally you would now click the button to Request Data Harvest. However, rather than wait for an administrator to trigger the harvest, you will trigger the harvesting yourself. Your user account for the workshop has been given sufficient privileges to trigger data harvests.

12. Trigger a harvest for your service.

a. Near the top of the Data Service Details web page, click **Harvest Data**.
b. Click **Begin Harvest**. In a moment, a link will appear to view the harvest log.
c. Click the Harvest log link to view progress. The file will indicate that the harvest is completed once it has completed. Refresh your browser to see any updates.

Now that the service is harvested, you can tag variables in the service to the ontology.

Tip

At this point, you may want to take a moment to stretch and relax. Sometimes it takes a few minutes for the results of the harvest to be committed to HIS Central's catalog database.

13. Click **My Data Services**.
14. Click **Details** for your service.
15. Click **List Variables**. This brings you to a page showing variables found in your service (Figure 0-2). If you don't see your variables here, then the harvesting was not completed.

Figure 0-2 List of variables for a service registered at HIS Central

16. Click **Map Variables**.

The HydroTagger opens, showing the CUAHSI HIS ontology. There are high level concepts in the center, with branches increasing in specificity as they move out to the leaf concepts. The variables that HIS Central found in your service are listed on the bottom left. Your mission here is to map your variables to the appropriate leaf concept. Users search upon these concepts, not whatever wacky variable names you used, when they search for data at HIS Central.

17. In your list of variables, click **Select** on the row for **chlorophyll a**. This adds that item to the Variable text box (Figure 0-3).
18. In the view of the ontology, click and drag with the mouse to locate the chlorophyll a concept leaf (Figure 0-3). It falls under **HydroSphere | Biological | Biological community | Pigment | Chlorophyll | Chlorophyll a**.
19. Double-click **Chlorophyll a** to add it to the Mapping text box at the bottom of the window (Figure 0-3).
20. Click **Map** (Figure 0-3).

Tip

If you can't find chlorophyll in the ontology, type chlorophyll in the Search box and click Search. Small red arrows will indicate keywords that match your search. Note that the Search box is only visible when the Variable and Mapping text boxes are not populated.

Figure 0-3 Mapping the chlorophyll variable

21. Repeat steps 17-20 for your other variables.
    a. **Temperature** – Hydrosphere | Physical | Temperature | Temperature, water
    b. **Nitrogen** – Hydrosphere | Chemical | Nutrient | Nitrogen | Nitrogen, total
    c. **Phosphorus** – Hydrosphere | Chemical | Nutrient | Phosphorus | Phosphorus, total
    d. **Total suspended solids** – Hydrosphere | Physical | Water | Water, suspended solids | Solids, total suspended

That's it! Your service is now registered and your variables are mapped. Let's take a look at how others will see your service when they navigate to HIS Central.

22. Click **Back** in your browser to navigate back to the variable list.
23. At the top of the web browser, click **All Data Services**.
24. Find your service in the list of services and click its Data Service Title. If you don't see your service on the first page, clink the link for subsequent pages at the bottom of the page. Or, click the CreatedDate field heading twice to sort the services by most recent creation date. Yours will be near the top of the list.

This brings you to the public details page for your service. Notice that the service details you entered are there, along with a map showing a box that defines the extent of your site locations (Figure 0-4).
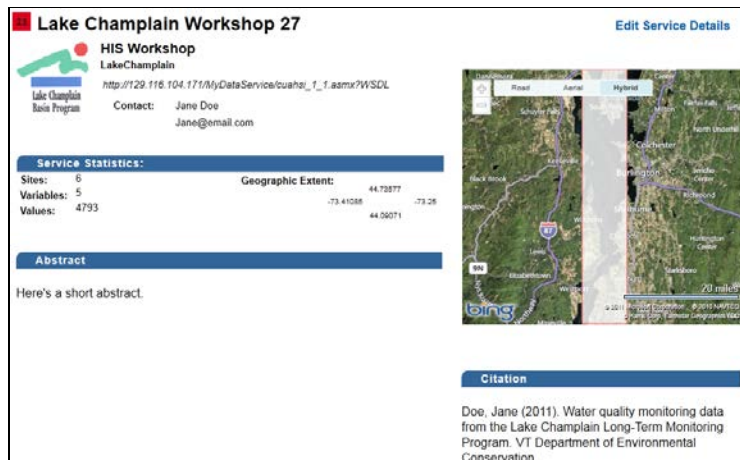
Figure 0-4 Public view of registered service at HIS Central (your images and details may vary)

Outstanding! Your service is now registered with HIS Central, and you've seen how others can navigate to HIS Central and discover your service. As a final check, let's find data from your service in HydroDesktop.

Viewing Your Data in HydroDesktop

The final step in the publication process is to check that your service is properly registered by viewing sites from your service in HydroDesktop. HydroDesktop is a desktop application for discovering, accessing, and analyzing time series data from WaterOneFlow services. Unlike HydroExcel, HydroDesktop doesn't require prior knowledge of your service URL before you can use it. Instead, HydroDesktop searches HIS Central's catalog for locations of time series of interest, giving you access to ALL publicly available services registered at HIS Central.

You'll use HydroDesktop to submit a query that should include sites from your Web service in the result. If you see your sites, then the data publication process is completed!

To search for your sites in HydroDesktop:
1. Open HydroDesktop and create a new North America project.

As you can see, HydroDesktop is a Geographic Information System (GIS) integrated with CUAHSI-HIS. Next you will tell HydroDesktop to use the sandbox HIS Central instead of the official version. Recall that you registered your service with the sandbox HIS Central (the one for testing and workshops) and not the official version.
2. In the ribbon, click the drop down arrow under the **Search** panel and then click **Advanced Settings** (Figure 0-5).

Figure 0-5 Accessing HydroDesktop advanced settings

3. In the Advanced Settings dialog, click **Custom** (Figure 0-6).
4. Enter the following URL into the box below Custom (Figure 0-6). This is the URL to the search services operated by the sandbox HIS Central in which you registered your service.
http://water.sdsc.edu/hiscentralsb/webservices/hiscentral.asmx
5. Click **OK** (Figure 0-6).

Figure 0-6 Setting the HIS Central URL

Now you will locate Lake Champlain. This is where your sites are located. An online basemap will be useful in finding the location.

6. In the ribbon, in the **Online Basemap** panel, select **ESRI World Topo** from the drop down list.
7. Using the **Zoom In** tool, zoom in to the northeastern United States to find Lake Champlain on the border with Canada between Vermont and New York.
8. In the Search panel, click the **Select by drawing Rectangle** tool (Figure 0-7).
9. Click and drag to draw a box around Lake Champlain (Figure 0-7).

A-55
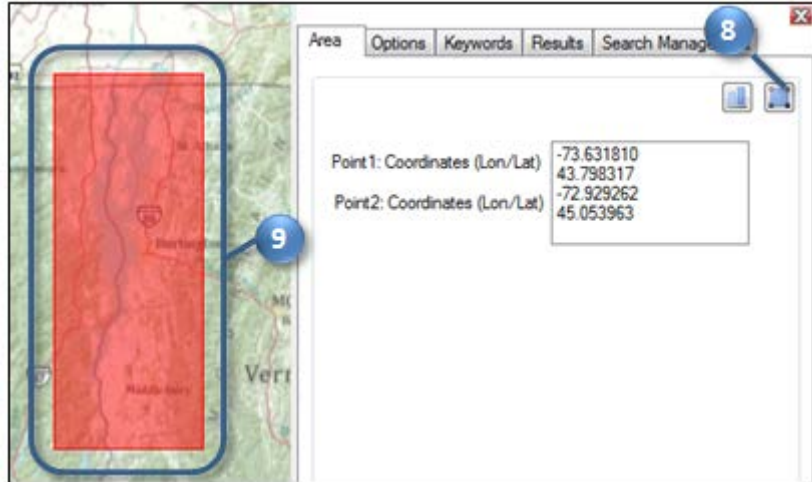
Figure 0-7 Specifying the area to search

10. In the **Search** panel, click the **Keywords** tab (Figure 0-8).
11. Type **nitrogen** in the keyword box (Figure 0-8). Keywords related to nitrogen are automatically selected in the keyword list.
12. Click **Nitrogen, total** (Figure 0-8). This is one of the variables included in your service.
13. Click the **Add** button (Figure 0-8). The nitrogen keyword is added to the list of keywords to search.
14. Click **Run Search** (Figure 0-8).
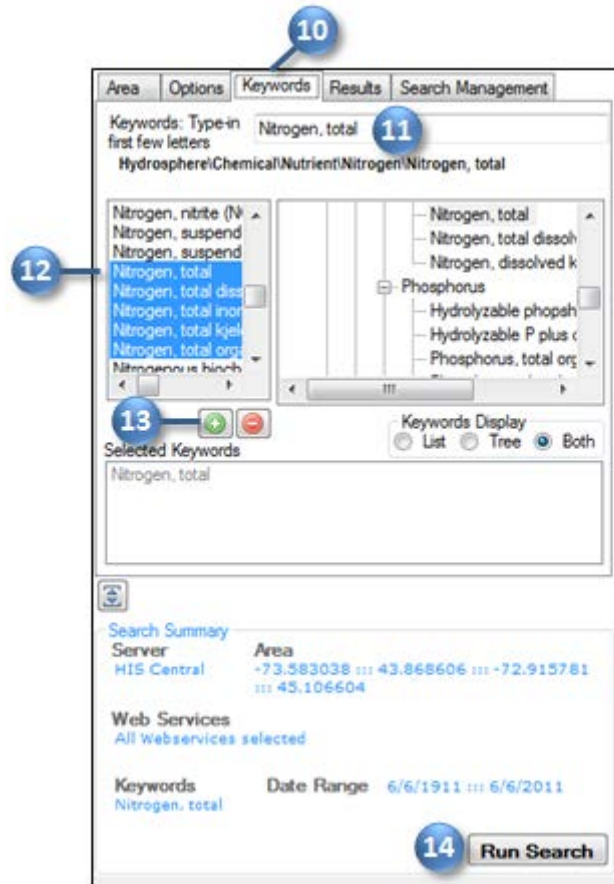
Figure 0-8 Specifying keywords to search

After a moment, search results are displayed in the map. The results include time series from the USGS National Water Information System as well as services registered as part of this workshop. If you see your sites, then you are finished! We already tested time series download using HydroExcel. But just in case you want to try the download in HydroDesktop, the steps are included below:

15. In the ribbon, click the **Select** tool to activate it.
16. Draw a box around one or more of your sites to select it for download.
17. In the Search panel, specify a theme name and Click Download Data. After a moment, the time series are downloaded and saved to your computer.
18. In the ribbon, click the Graph tab and select a time series to plot a graph. If everything worked, you should now be looking at a graph of your data. You discovered and accessed the data just like other users can, meaning you have achieved success in publishing your data online!

Congratulations! You have now completed the HIS data publication process. You started from raw data, loaded the data into an ODM database, wrapped Web services around the database, and registered the Web service at a sandbox version of HIS Central. Had you been using the official HIS Central, these data would now be participating in the national system along with data from the USGS, EPA, and a number of other agencies and academic institutions!

If you'd like to know more about the data publication process, please visit the HIS website at http://his.cuahsi.org. You can check for new releases of helpful HIS tools, and even partner with the HIS

team in developing the next generation of HIS software and methods to further hydrologic information science.

This concludes the workshop.

Appendix A: Uninstallation Instructions

If you want to erase your footsteps from this workshop, here's how.

HIS Central

Only administrators can permanently delete services at HIS Central. The workshop instructor will delete all services registered during the workshop within a few weeks after the workshop has concluded. If you register another service in the future and would like it permanently deleted, feel free to email the workshop instructor who will tell the HIS Central administrator to delete the service for you.

In the meantime, you can hide the service so that others cannot see it.

To hide a service at HIS Central:
1. Log in to **HIS Central**.
2. Click **My Data Services**.
3. Click **Details** for the service that you want to hide.
4. Click **Edit Details**.
5. Uncheck **Is service public**.
6. Click **Update**.

WaterOneFlow

To uninstall your WaterOneFlow Web service from your server, you will remove it from IIS and then delete the files on your computer.

To uninstall WaterOneFlow:
1. Open **IIS**.
2. Expand the tree view on the left to find **MyDataService**.
3. Right-click your service and click **Remove**. Click Yes if prompted to remove the application.
4. Optional: If you have no other services using the WaterOneFlow application pool, then in the tree view, click **Application Pools**, find **WaterOneFlow**, and click to **Remove** the WaterOneFlow application pool.
5. Close IIS.
6. Open **Windows Explorer**.
7. Navigate to your **MyDataService** folder.

Deleting the files is a two-step process. First you delete the service files since they have a lock on supporting files. Then you delete the supporting files.

To delete WaterOneFlow files:
1. In the MyDataService folder, **delete all ASP.NET Server Page files**. These are the files with the .aspx or .asmx extension. (Hint: Sort by Type.)
2. Delete the **MyDataService** folder.

Streaming Data Loader Scheduled Task

To remove the scheduled task for the Streaming Data Loader:
1. Open **Task Scheduler**.

2. In the list of **Active Tasks**, double-click **Streaming Data Loader**.
3. Click **Actions | Delete**. Click **Yes** to confirm. Close Task Scheduler.

ODM Database

To clean up your instance of SQL Server:
1. Log into **SQL Management Studio**.
2. In the Object Explorer, find your **MyWaterData** database.
3. Right-click **MyWaterData** and click **Delete**. Click **OK** to delete the database. This both detaches your database and deletes the files from your computer.
4. Optional: If you have no other databases associated with your webclient SQL Server login, then you can delete it.
    a. In the Object Explorer, expand **Security | Logins** to find your webclient login.
    b. Right-click **webclient** and click **Delete**.   Click **OK** to delete the login.
5. Repeat step 4 for any other ODM-related logins that you created.

CUAHSI-HIS and Related Software

To uninstall HIS software:
1. Open **Programs and Features**.
2. Uninstall the following software:
    a. CUAHSI HydroObjects
    b. Google Earth
    c. HydroDesktop
    d. Microsoft SQL Server or SQL Server Express (please see your system administrator first since this program would take a long time to reinstall if desired)
    e. ODM Tools
    f. ODM Data Loader
    g. ODM Streaming Data Loader
    h. Optional: Microsoft Office (please see your system administrator first since this program is not free)
3. Close Programs and Features.
4. To uninstall **HydroExcel**, delete the spreadsheet file (**HydroExcel115.xlsb**).
5. Delete the HydroExcel115 folder, located in the same directory as the HydroExcel spreadsheet file. This folder contains KML generated during the Google Earth portion of the exercise.

This page intentionally left blank.

CUAHSI Community Observations Data Model (ODM) Version 1.1 Design Specifications

May 2008

David G. Tarboton[1], Jeffery S. Horsburgh[1], David R. Maidment[2]

Abstract

The CUAHSI Hydrologic Information System project is developing information technology infrastructure to support hydrologic science. One aspect of this is a data model for the storage and retrieval of hydrologic observations in a relational database. The purpose for such a database is to store hydrologic observations data in a system designed to optimize data retrieval for integrated analysis of information collected by multiple investigators. It is intended to provide a standard format to aid in the effective sharing of information between investigators and to allow analysis of information from disparate sources both within a single study area or hydrologic observatory and across hydrologic observatories and regions. The observations data model is designed to store hydrologic observations and sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used. A relational database format is used to provide querying capability to allow data retrieval supporting diverse analyses. A generic template for the observations database is presented. This is referred to as the Observations Data Model (ODM).

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) is an organization representing more than 100 universities and is sponsored by the National Science Foundation to provide infrastructure and services to advance the development of hydrologic science and education in the United States. The CUAHSI Hydrologic Information System (HIS) is being developed as a geographically distributed network of hydrologic data sources and functions that are integrated using web services so that they function as a connected whole. One aspect of the CUAHSI HIS is the development of a standard database schema for use in the storage of point observations in a relational database. This is referred to as the point Observations Data Model (ODM) and is intended to allow for comprehensive analysis of information collected by multiple investigators for varying purposes. It is intended to expand the ability for data analysis by providing a standard format to share data among investigators and to facilitate analysis of information from disparate sources both within a single study area or hydrologic observatory and across hydrologic observatories and regions. The ODM is designed to store hydrologic observations with sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used. Although designed specifically with hydrologic observation data in mind, this data model has a simple and general structure that will also accommodate a wide range of other data, such as from other environmental observatories or observing networks.

---

[1] Utah Water Research Laboratory, Utah State University
[2] Center for Research in Water Resources, University of Texas at Austin

ODM uses a relational database format to allow for ease in querying and data retrieval in support of a diverse range of analyses. Reliance on databases and tables within databases also provides the capability to have the model scalable from the observations of a single investigator in a single project through the multiple investigator communities associated with a hydrologic observatory and ultimately to the entire set of observations available to the CUAHSI community. ODM is focused on observations made at a point. A relational database model with individual observations recorded as individual records (an atomic model) has been chosen to provide maximum flexibility in data analysis through the ability to query and select individual observation records. This approach carries the burden of record level metadata, so it is not appropriate for all variables that might be observed. For example, individual pixel values in large remotely sensed images or grids are inappropriate for this model.

This data model is presented as a generic template for a point observations database, without reference to the specific implementation in a database management system. This is done so that the general design is not limited to any specific proprietary software, although we expect that implementations will take advantage of capabilities of specific software. It should be possible to implement ODM in a variety of relational database management systems, or even in a set of text tables or variable arrays in a computer program. However, to take full advantage of the relationships between data elements, the querying capability of a relational database system is required. By presenting the design at a general conceptual level, we also avoid implementation specific detail on the format of how information is represented. See the discussion of Dates and Times under ODM features below for an example of the distinction between general concepts and implementation specific details.

Version Information

ODM has evolved from an initial design presented at a CUAHSI workshop held in Austin during March, 2005 (Maidment, 2005) that was then widely reviewed with comments being received from 22 individuals (Tarboton, 2005). These reviews served as the basis for a redesign that was presented at a CUAHSI workshop in Duke during July, 2005 and presented as part of the CUAHSI HIS status report (Horsburgh et al., 2005). Following this presentation of the design, the data model was reviewed and commented on by a number of others, including the CLEANER (Collaborative Large-scale Engineering Analysis Network for Environmental Research) cyberinfrastructure committee. Further versions of the Observations Data Model were circulated in April, June and October 2006. These documented changes made in the evolution of this design. The fundamental design, however, has not changed since the status report presentation of the model (Horsburgh et al., 2005) but many table and field names have been changed. Tables have also been added to give spatial reference information, metadata information, and to define controlled vocabularies. Version 1.0 of ODM, which was the first release version of ODM, has been implemented and tested within the WATERS network of test bed sites and was documented in Water Resources Research (Horsburgh et al., 2008). This document describes the second release version of the data model design, which has been named ODM Release Version 1.1, and has been so named to correspond to the Version 1.1 release of the CUAHSI HIS. This document supersedes the previous documents.

In general, the following changes have been made for Version 1.1:

- All integer IDs serving as the primary key for tables in ODM have been changed to auto number/identity fields.
- Text field lengths have been relaxed in some cases and have been standardized according to the following scheme: codes = 50 characters, terms = 255 characters, links = 500 characters, definitions/explanations = unlimited.
- Check constraints have been defined for the Latitude and Longitude fields in the Sites table.
- Check constraints have been added to many of the fields in ODM to constrain the characters that are valid for those fields (see Appendix A for details).
- Relationships have been added between controlled vocabulary tables and the tables that contain the fields that they define. This was done to more rigorously enforce the ODM controlled vocabularies.
- Unique constraints were placed on both SiteCode in the Sites table and VariableCode in the Variables table.
- The controlled vocabulary was relaxed on the QualityControlLevels table to allow more detailed versioning of data series. A QualityControlLevelCode was also added to this table to facilitate this.
- A Citation field was added to the Sources table to provide a place for a formal citation for data in the database.
- A Speciation field was added to the Variables table. This provides a place to store information about the speciation of chemistry observations. A SpeciationCV controlled vocabulary table was added to define this field.
- An ODMVersion table was added to store the version number of the database.
- The SeriesCatalog table has been updated based on the addition of the above fields.

Hydrologic Observations

Many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels, and precipitation. National databases such as USGS' National Water Information System (NWIS) and USEPA's data Storage and Retrieval (STORET) system contain a wealth of data, but, in general, these national data repositories have different data formats, storage, and retrieval systems, and combining data from disparate sources can be difficult. The problem is compounded when multiple investigators are involved (as would be the case at proposed CUAHSI Hydrologic Observatories) because everyone has their own way of storing and manipulating observational data. There is a need within the hydrologic community for an observations database structure that presents observations from many different sources and of many different types in a consistent format.

Hydrologic observations are identified by the following fundamental characteristics:

- The location at which the observations were made (space)
- The date and time at which the observations were made (time)
- The type of variable that was observed, such as streamflow, water surface elevation, water quality concentration, etc. (variable)

These three fundamental characteristics may be represented as a data cube (Figure 1), where a particular observed data value (D) is located as a function of where it was observed (L), its time of observation (T), and what kind of variable it is (V), thus forming D(L,T,V).
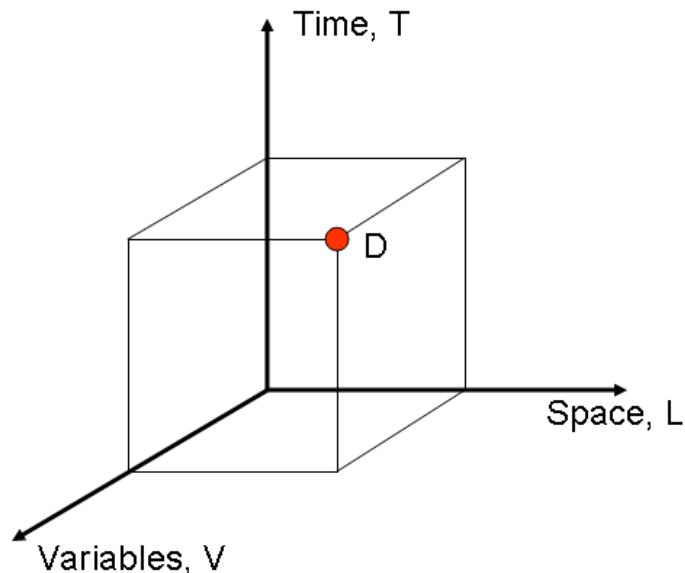
Figure 1. A measured data value (D) is indexed by its spatial location (L), its time of measurement (T), and what kind of variable it is (V).

In addition to these fundamental characteristics, there are many other distinguishing attributes that accompany observational data. Many of these secondary attributes provide more information about the three fundamental characteristics mentioned above. For example, the location of an observation can be expressed as a text string (i.e., "Bear River Near Logan, UT"), or as latitude and longitude coordinates that accurately delineate the location of the observation. Other attributes can provide important context in interpreting the observational data. These include data qualifying comments and information about the organization that collected the data. The fundamental design decisions associated with the ODM involve

choices as to how much supporting information to include in the database and whether to store (and potentially repeat) this information with each observation or save this information in separate tables with key fields used to logically associate observation records with the associated information in the ancillary tables. Table 1 presents the general attributes associated with a point observation that we judged should be included in the generic ODM design.

Table 1. ODM attributes associated with an observation

Attribute	Definition
Data Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
Date and Time	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
Variable Name	The name of the physical, chemical, or biological quantity that the data value represents (e.g. streamflow, precipitation, temperature)
Speciation	For concentration measurements, the species in which the concentration is expressed (e.g., as N, or as NO_3, or as NH_4)
Location	The location at which the observation was made (e.g. latitude and longitude)
Units	The units (e.g. m or m3/s) and unit type (e.g. length or volume/time) associated with the variable
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method and whether the observations are regularly recorded on that interval
Offset	Distance from a reference point to the location at which the observation was made (e.g. 5 meters below water surface)
Offset Type/ Reference Point	The reference point from which the offset to the measurement location was measured (e.g. water surface, stream bank, snow surface)
Data Type	An indication of the kind of quantity being measured (e.g. a continuous, minimum, maximum, or cumulative measurement)
Organization	The organization or entity providing the measurement
Censoring	An indication of whether the observation is censored or not
Data Qualifying Comments	Comments accompanying the data that can affect the way the data is used or interpreted (e.g. holding time exceeded, sample contaminated, provisional data subject to change, etc.)
Analysis Procedure/ Method	An indication of what method was used to collect the observation (e.g. dissolved oxygen by field probe or dissolved oxygen by Winkler Titration) including quality control and assurance that it has been subject to
Source	Information on the original source of the observation (e.g. from a specific organization, agency, or investigator 3rd party database)
Sample Medium	The medium in which the sample was collected (e.g. water, air, sediment, etc.)
Value Category	An indication of whether the data value represents an actual measurement, a calculated value, or is the result of a model simulation

Observations Data Model

The schema of the Observations Data Model is given in Figure 2. Appendix A gives details of each table and each field in this generic data model schema. Appendix A serves as the data dictionary for the data model and documents specific database constraints, data types, examples, and best practices. The primary

table that stores point observation values is the DataValues table at the center of the schema in Figure 2. Logical relationships between fields in the data model are shown and serve to establish the connectivity between the observation values and associated ancillary information. Details of the relationships are given in Table 2. Figure 2 shows each of the controlled vocabulary tables and their relationships to the table containing the field that they define. Controlled vocabulary tables are highlighted with red headers. In Figure 2, each of the mandatory fields is shown in bold text, whereas optional fields are shown in regular text.

Monitoring Site Locations

SpatialReferences
- (M) SpatialReferenceID {PK}
- (O) SRSID
- (M) SRSName
- (O) IsGeographic
- (O) Notes

Sites
- (M) SiteID {PK}
- (M) SiteCode
- (M) SiteName
- (M) Latitude
- (M) Longitude
- (M) LatLongDatumID {FK}
- (O) Elevation_m
- (O) VerticalDatum
- (O) LocalX
- (O) LocalY
- (O) LocalProjectionID {FK}
- (O) PosAccuracy_m
- (O) State
- (O) County
- (O) Comments

RefLatLong ▶ 1..1   0..*

ReferenceLocalCoords ▶   0..*

DefinesCV ▶   0..* 1..1

VerticalDatumCV
- (M) Term
- (O) Definition

Diagram Key

Table
- (M) Field cannot be null
- (O) Field can be Null

Controlled Vocabulary
- (M) Field cannot be null
- (O) Field can be Null

Cardinality
- 1..1 One and only one
- 1..* One or many
- 0..* Zero or many
- 0..1 Zero or one

ODM Version

ODMVersion
- (M) Version Number

Data Sources

Sources
- (M) SourceID {PK}
- (M) Organization
- (M) SourceDescription
- (O) SourceLink
- (M) ContactName
- (M) Phone
- (M) Email
- (M) Address
- (M) City
- (M) State
- (M) ZipCode
- (M) Citation
- (M) MetadataID {FK}

DescribeSource ▶ 1..1   0..*

ISOMetadata
- (M) MetadataID {PK}
- (M) TopicCategory
- (M) Title
- (M) Abstract
- (M) ProfileVersion
- (O) MetadataLink

DefinesCV ▶ 0..* 1..1

TopicCategoryCV
- (M) Term
- (O) Definition

Observation Values

◀ Generate

◀ Create

Locate ▶   0..* 0..*

DataValues
- (M) ValueID {PK}
- (M) DataValue
- (O) ValueAccuracy
- (M) LocalDateTime
- (M) UTCOffset
- (M) DateTimeUTC
- (M) SiteID {FK}
- (M) VariableID {FK}
- (O) OffsetValue
- (O) OffsetTypeID {FK}
- (M) CensorCode
- (O) QualifierID {FK}
- (M) MethodID {FK}
- (M) SourceID {FK}
- (O) SampleID {FK}
- (O) DerivedFromID
- (M) QualityControlLevelID {FK}

◀ ResultIn   0..*

DefinesCV ▶ 0..*

1..*   1..1

CensorCodeCV
- (M) Term
- (O) Definition

AreDerivedFrom ▶   0..*

DerivedFrom
- (M) DerivedFromID {FK}
- (M) ValueID {FK}

◀ Group

Data Collection Methods

Methods
- (M) MethodID {PK}
- (M) MethodDescription
- (O) MethodLink

1..1

LabMethods
- (M) LabMethodID {PK}
- (M) LabName
- (M) LabOrganization
- (M) LabMethodName
- (M) LabMethodDescription
- (O) LabMethodLink

1..1

Samples
- (M) SampleID {PK}
- (M) SampleType
- (M) LabSampleCode
- (M) LabMethodID {FK}

◀ Analyze   0..*   0..1

◀ DefinesCV   0..* 1..1

SampleTypeCV
- (M) Term
- (O) Definition

Units

Units
- (M) UnitsID {PK}
- (M) UnitsName
- (M) UnitsType
- (M) UnitsAbbreviation

DescribeOffsetUnits ▶ 1..1

DescribeTimeUnits ▶

DescribeVariableUnits ▶

Offsets

OffsetTypes
- (M) OffsetTypeID {PK}
- (M) OffsetUnitsID {FK}
- (M) OffsetDescription

OffsetLocationOf ▶ 0..* 0..1

◀ Characterize

Variables

VariableNameCV
- (M) Term
- (O) Definition

DefinesCV ▶ 1..1 0..*

SpeciationCV
- (M) Term
- (O) Definition

DefinesCV ▶ 1..1 0..*

SampleMediumCV
- (M) Term
- (O) Definition

DefinesCV ▶ 0..*

ValueTypeCV
- (M) Term
- (O) Definition

DefinesCV ▶ 0..*

DataTypeCV
- (M) Term
- (O) Definition

DefinesCV ▶ 1..1 0..*

GerneralCategoryCV
- (M) Term
- (O) Definition

DefinesCV ▶ 1..1

Variables
- (M) VariableID {PK}
- (M) VariableCode
- (M) VariableName
- (M) Speciation
- (M) VariableUnitsID {FK}
- (M) SampleMedium
- (M) ValueType
- (M) IsRegular
- (M) TimeSupport
- (M) TimeUnitsID {FK}
- (M) DataType
- (M) GeneralCategory
- (M) NoDataValue

1..1

◀ Represent

CharacterizeCategorical ▶

Categorical Data

Categories
- (M) VariableID {FK}
- (M) DataValue
- (M) CategoryDescription

0..1   0..*

◀ SpecifyQualityOf

Data Qualifiers

◀ Qualify

QualityControlLevels
- (M) QualityControlLevelID {PK}
- (M) QualityControlLevelCode
- (M) Definition
- (M) Explanation

1..1

Qualifiers
- (M) QualifierID {PK}
- (O) QualifierCode
- (M) QualifierDescription

0..1

Value Grouping

Groups
- (M) GroupID {FK}
- (M) ValueID {FK}

◀ Describe   0..* 0..*

GroupDescriptions
- (M) GroupID {PK}
- (O) GroupDescription

1..1

SeriesCatalog
- (P) SeriesID {PK}
- (P) SiteID {FK}
- (P) SiteCode
- (P) SiteName
- (P) VariableID {FK}
- (P) VariableCode
- (P) VariableName
- (P) Speciation
- (P) VariableUnitsID {FK}
- (P) VariableUnitsName
- (P) SampleMedium
- (P) ValueType
- (P) TimeSupport
- (P) TimeUnitsID {FK}
- (P) TimeUnitsName
- (P) DataType
- (P) GeneralCategory
- (P) MethodID {FK}
- (P) MethodDescription
- (P) SourceID {FK}
- (P) Organization
- (P) SourceDescription
- (P) Citation
- (P) QualityControlLevelID {FK}
- (P) QualityControlLevelCode
- (P) BeginDateTime
- (P) EndDateTime
- (P) BeginDateTimeUTC
- (P) EndDateTimeUTC
- (P) ValueCount

Figure 2. Observations Data Model schema.

Table 2. Observations Data Model Logical Relationships

Relationships that define <u>ancillary information</u> about data values

Table	Field	Type	Field	Table
DataValues	SiteID	* <-> 1	SiteID	Sites
DataValues	VariableID	* <-> 1	VariableID	Variables
DataValues	OffsetTypeID	* <-> 1	OffsetTypeID	OffsetTypes
DataValues	QualifierID	* <-> 1	QualifierID	Qualifiers
DataValues	MethodID	* <-> 1	MethodID	Methods
DataValues	SourceID	* <-> 1	SourceID	Sources
DataValues	SampleID	* <-> 1	SampleID	Samples
DataValues	QualityControlLevelID	* <-> 1	QualityControlLevelID	QualityControlLevels

Relationships that define <u>derived from groups</u>

Table	Field	Type	Field	Table
DataValues	DerivedFromID	* <-> *	DerivedFromID	DerivedFrom
DataValues	ValueID	1 <-> *	ValueID	DerivedFrom

Relationships that define <u>groups</u>

Table	Field	Type	Field	Table
DataValues	ValueID	1 <-> *	ValueID	Groups
GroupDescriptions	GroupID	1 <-> *	GroupID	Groups

Relationships used to define <u>categories for categorical data</u>

Table	Field	Type	Field	Table
Variables	VariableID	1 <-> *	VariableID	Categories
DataValues	DataValue	* <-> 1	DataValue	Categories

Relationships used to define the <u>Units</u>

Table	Field	Type	Field	Table
Units	UnitsID	1<->*	VariableUnitsID	Variables
Units	UnitsID	1<->*	TimeUnitsID	Variables
Units	UnitsID	1<->*	OffsetUnitsID	OffsetTypes

Relationship used to define the <u>Sample Laboratory Methods</u>

Table	Field	Type	Field	Table
LabMethods	LabMethodID	1<->*	LabMethodID	Samples

Relationships used to define the <u>Spatial References</u>

Table	Field	Type	Field	Table
SpatialReferences	SpatialReferenceID	1<->*	LatLongDatumID	Sites
SpatialReferences	SpatialReferenceID	1<->*	LocalProjectionID	Sites

Relationship used to define the <u>ISOMetaData</u>

Table	Field	Type	Field	Table
IsoMetaData	MetadataID	1<->*	Sources	MetadataID

Relationships used to define <u>Controlled Vocabularies</u>

Table	Field	Type	Field	Table
VerticalDatumCV	Term	1<->*	VerticalDatum	Sites
SampleTypeCV	Term	1<->*	SampleType	Samples
VariableNameCV	Term	1<->*	VariableName	Variables
ValueTypeCV	Term	1<->*	ValueType	Variables
DataTypeCV	Term	1<->*	DataType	Variables
SampleMediumCV	Term	1<->*	SampleMedium	Variables
SpeciationCV	Term	1<->*	Speciation	Variables
GeneralCategoryCV	Term	1<->*	GeneralCategory	Variables
TopicCategoryCV	Term	1<->*	TopicCategory	ISOMetadata
CensorCodeCV	Term	1<->*	CensorCode	DataValues

Relationship type is indicated as One to One (1<->1), One to Many (1<->*), Many to One (*<->1) and Many to Many (*<->*). The first set of relationships defines the links to tables that contain ancillary information. They are used so that only compact (integer) identifiers are stored with each data value and thus repeated many times while the more voluminous ancillary information is stored to the side and not repeated. The second set of relationships defines derived from groupings used to specify data values that have been used to derive other data values. The third set of relationships defines logical groupings of data values. The fourth set of relationships is used to specify the categories associated with categorical variables. The fifth set of relationships is used to define the units. The sixth set of relationships associates laboratory methods with samples. The seventh set of relationships associates sites with the Spatial Reference System used to define the location. The eigth set of relationships associates project and dataset level metadata with each data source. The last set of relationships defines the linkage between the controlled vocabulary fields and the tables that stored the acceptable terms for those fields. Details of how these relationships work are given in the discussion of features of the data model design below.

Features of the Observations Data Model Design

Geography
ODM is intended to be independent of the geographical representation of the site locations. The geographic location of sites is specified through the Latitude, Longitude, and Elevation information in the Sites table, and optionally local coordinates, which may be in a standard geographic projection for the study area or in a locally defined coordinate system specific to a study area. Each site also has a unique identifier, SiteID, which can be logically linked to one or more objects in a Geographic Information System (GIS) data model. For example, Figure 3 depicts a one-to-one relationship between sites within ODM and HydroPoints within the Arc Hydro Framework Data Model (Maidment, 2002) used to represent objects in a digital watershed. In simple implementations, SiteID may have the same integer value as the identifier for the associated GIS object, HydroID in this case. In more complex implementations, and especially when multiple databases are merged into a single ODM, it may not be possible to preserve the simple one-to-one relationship between SiteID and HydroID with each of these fields holding the same integer identifier values. In these cases, where SiteID and HydroID are not the same, a coupling table would be used to associate the ODM SiteIDs used to identify sites with HydroIDs in the Arc Hydro data model.

SiteID must be unique within an instance of ODM. This could, for example, be achieved by assigning SiteIDs from a master table. The linkage between SiteIDs and GIS object IDs is intended to be generic and suitable for use with any geographic data model that includes information specifying the location of sites. For example, a linear referencing system on a river network, such as the National Hydrography

Dataset, might be used to specify the location of a site on a river network. Addressing relative to specific hydrologic objects through the SiteID field provides direct and specific location information necessary for proper interpretation of data values. Information from direct addressing relative to hydrologic objects is often of greater value to a user than the simple Latitude and Longitude information stored in the ODM Sites table. For example, it is more useful to know that a stream gage is on such and such a stream rather than simply its latitude and longitude.
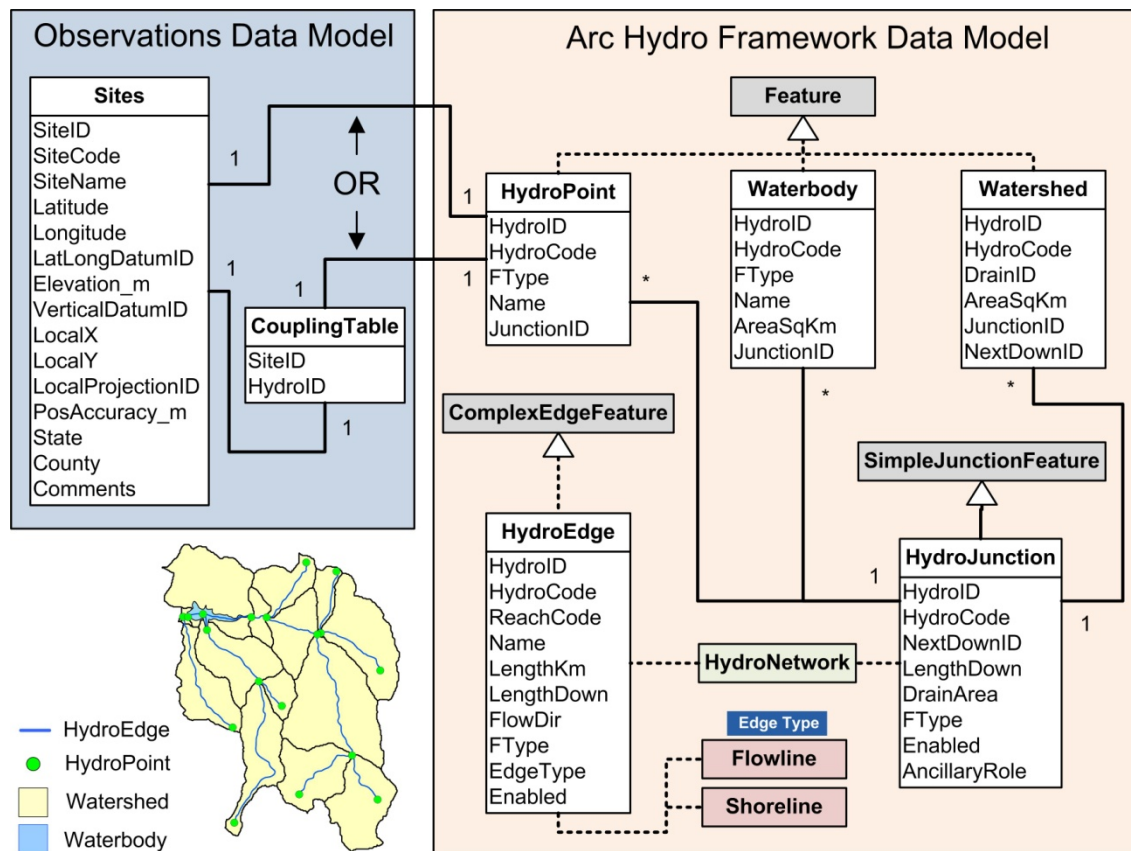
Figure 3. Arc Hydro Framework Data Model and Observations Data Model related through SiteID field in the Sites table.

Series Catalog

A "data series" is an organizing principle used in ODM. A data series consists of all the data values associated with a unique site, variable, method, source, and quality control level combination in the DataValues table. The SeriesCatalog table lists data series, identifying each by a unique series identifier, SeriesID. This table is essentially a summary of many of the tables in the ODM and is not required to maintain the integrity of the data. However, it serves to provide a listing of all the distinct series of data values of a specific variable at a specific site. By doing so, this table provides a means by which users can execute most common data discovery queries (i.e., which variables have data at a site, etc.) without the overhead of querying the entire DataValues table, which can become quite large.

The SeriesCatalog table is also intended to support CUAHSI Web Service method queries such as GetSiteInfo, which returns information about a monitoring site within an instance of the ODM including the variables that have been measured at that site. It should be noted that data series, as they are defined here, do not distinguish between different series of the same variable at the same site but measured with different offsets. If for example temperature was measured at two different offsets by two different sensors at one site, both sets of data would fall into one data series for the purposes of the SeriesCatalog

table. In these cases, interpretation or analysis software will need to specifically examine and parse the offsets by examining the offset associated with each data value. The SeriesCatalog table does not do this because its principal purpose is data discovery, which we did not want to be overly complicated. The SeriesCatalog table should be programmatically generated and modified as data are added to the database.

Accuracy

Each data value in the DataValues table has an associated attribute called ValueAccuracy. This is a numeric value that quantifies the total measurement accuracy defined as the nearness of a measurement to the true or standard value. Since the true value is not known, the ValueAccuracy is estimated based on knowledge of the instrument accuracy, measurement method, and operational environment. The ValueAccuracy, which is also called the uncertainty of the measurement, compounds the estimates of both bias and precision errors. Bias errors are generally fixed or systematic and cannot be determined statistically, while precision errors are random, being generated by the variability in the measurement system and operational environment. Figure 4 illustrates the effects of these errors on a sample of measurements. Bias errors are usually estimated through specially designed experiments (calibrations). The precision errors are determined using statistical analysis by quantifying the measurement scatter, which is proportional to the standard deviation of the sample of repeated measurements. The total error is obtained by the root-sum-square of the estimates for bias and precision errors involved in the measurement. Figure 5 gives another illustration of the ValueAccuracy concept based on the analogy of a target, where the bulls eye at the center represents the true value.

ValueAccuracy is a data value level attribute because it can change with each measurement, dependent on the instrument or measurement protocol. For example, if streamflow is measured using a V-notch weir, it is actually the stage that is measured, with accuracy limited by the precision and bias of the depth recording instrument. The conversion to discharge through the stage-discharge relationship results in greater absolute error for larger discharges. Inclusion of the ValueAccuracy attribute, which will be blank for many historic datasets because historically accuracy has not been recorded, adds to the size of data in the ODM, but provides a way for factoring the accuracy associated with measurements into data analysis and interpretation, a practice that should be encouraged.
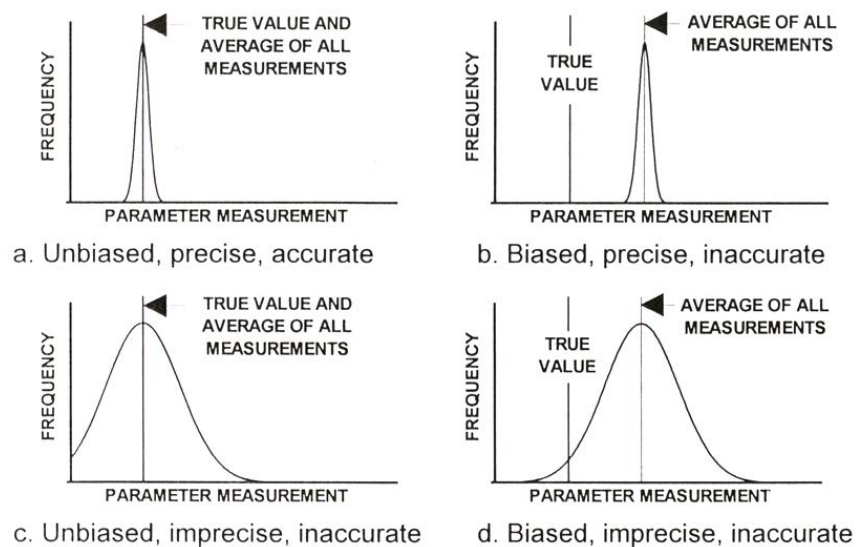
Figure 4. Illustration of measurement error effect (Source: AIAA, 1995).
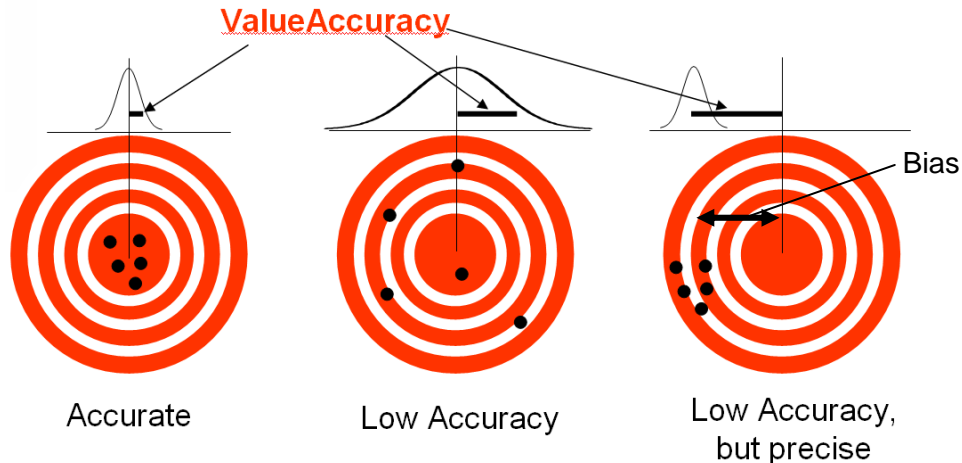
Figure 5. Illustration of Accuracy versus Precision (adapted from Wikipedia http://en.wikipedia.org/wiki/Accuracy).

In designing ODM, consideration was given to the suggestion by some reviewers to record bias and precision separately, in addition to ValueAccuracy for each data value. This has not been done at this release in the interest of parsimony and also because quantifying these separate components of the error is difficult. We suggest that for most measurements there should be the presumption that they are unbiased and that ValueAccuracy quantifies the precision and accuracy in the judgment of the investigator responsible for collecting the data. For cases where there is specific bias and precision information to complement the ValueAccuracy attribute, this could be recorded in the ODM as a separate variable, e.g. discharge precision, or temperature bias. The groups and derived from features (see below) could be used to associate these variables with their related observations. For measurements that are known to be biased, we suggest that the bias could be quantified by other reference measurements that should also be placed in the database and that a new set of corrected measurements that have had the bias removed should be added to the database at a higher quality control level. These new measurements should have a lower ValueAccuracy value to reflect the improvement in accuracy by removal of the bias. The method and derived from information for these corrected measurements should give the bias removal method and refer to the data used to quantify and remove the bias.

Offset
Each record in the DataValues table has two optional fields OffsetValue and OffsetTypeID. These are used to record the location of an observation relative to an appropriate datum, such as "depth below the water surface" or "depth below or above the ground." The OffsetTypeID references an OffsetValue into an OffsetTypes table that gives units and definition associated with the OffsetValue. This design only has the capability to represent one offset for each data value. In cases (which we expect to be rare) when there are multiple offsets (e.g. distance in from a stream bank and depth below the surface) one of the offsets will need to be distinguished as a separate variable.

Spatial Reference and Positional Accuracy
Unambiguous specification of the location of an observation site requires that the horizontal and vertical datum used for latitude, longitude, and elevation be specified. The SpatialReferences table is provided for this purpose to record the name and EPSG code of each Spatial Reference System used. EPSG codes are numeric codes associated with coordinate system definitions published by the OGP Surveying and Positioning Committee (http://www.epsg.org/). A non-standard Spatial Reference System, such as, for example, a local grid at an experimental watershed, may be defined in the SpatialReferences table Notes field. The accuracy with which the location of a monitoring site is known is quantified using the PosAccuracy_m field in the Sites table. This is a numeric value intended to specify the uncertainty (as a

standard deviation or root mean square error) in the spatial location information (latitude and longitude or local coordinates) in meters. Using a large number for PosAccuracy_m (e.g. 2000 m) accommodates entry of data collected for a study area where the precise location where the observation was recorded is not known.

Groups and Derived from Associations

The DerivedFrom and Groups tables fulfill the function of grouping data values for different purposes. These are tables where the same identifier (DerivedFromID or GroupID) can appear multiple times in the table associated with different ValueIDs, thereby defining an associated group of records. In the DerivedFrom table this is the sole purpose of the table, and each group so defined is associated with a record in the DataValues table (through the DerivedFromID field in that table). This record would have been derived from the data values identified by the group. The method of derivation would be given through the methods table associated with the data value. This construct is useful, for example, to identify the 96 15-minute unit streamflow values that go into the estimate of the mean daily streamflow. Note that there is no limit to how many groups a data value may be associated with, and data values that are derived from other data values may themselves belong to groups used to derive other data values (e.g. the daily minimum flow over a month derived from daily values derived from 15 minute unit values). Note also that a derived from group may have as few as one data value for the case where a data value is derived from a single more primitive data value (e.g., discharge from stage). Through this construct the ODM has the capability to store raw observation values and information derived from raw observations, while preserving the connection of each data value to its more primitive raw measurement.

The GroupID relationship that appears in Table 2 is designated as one-to-many because there will be many records in the Groups table that have the same GroupID, but different ValueID, that serve to define the group. In Figure 1, the Group relationship is labeled 1..*, at the DataValues table and 0..* at the Groups table. This indicates that a group may comprise one or more data values and that a data value may be included in 0 or more groups. Similarly, there will be many records in the DerivedFrom table that have the same DerivedFromID, but different ValueID that serve to define the group of data values from which a data value is derived. Logically a data value should not be in a DerivedFrom group upon which it is derived from. If this can be programmatically checked by the system, then this sort of circularity error could be prevented.

The method description in the Methods table associated with a data value that has a DerivedFromID should describe the method used for deriving the particular data value from other data values (e.g. calculating discharge from a number of velocity measurements across a stream). The relationship between the DataValues table DerivedFromID field and DerivedFrom table DerivedFromID field is many-to-many (*<->*) because it can occur that the same group of data values is used to derive more than one derived data value. In Figure 1, the AreDerivedFrom relationship between the data values and DerivedFrom table actually depicts both relationships between these tables listed in Table 2. The AreDerivedFrom relationship is labeled 1..* at the DataValues table and 0..* at the DerivedFrom table to indicate that a derived from group may comprise 1 or more data values and that a data value may be a member of 0 or more derived from groups.

Dates and Times

Unambiguous interpretation of date and time information requires specification of the time zone or offset from universal time (UTC). A UTCOffset field is included in the DataValues table to ensure that local times recorded in the database can be referenced to standard time and to enable comparison of results across databases that may store data values collected in different time zones (e.g. compare data values from one hydrologic observatory to those collected at another hydrologic observatory located across the country). A design choice here was to have UTCOffset as a record level qualifier because even though the time zone, and hence offset, is likely the same for all measurements at a site, the offset may change

due to daylight savings. Some investigators may run data loggers on UTC time, while others may use local time adjusting for daylight saving time. To avoid the necessity to keep track of the system used, or impose a system that might be cumbersome and lead to errors, we decided that if the offset was always recorded, the precise time would be unambiguous and would reduce the chance for interpretation errors. A field DateTimeUTC is also included as a record level attribute associated with each data value. This provides a consistent time for querying and sorting data values. There is a level of redundancy between LocalDateTime, UTCOffset and DateTimeUTC. Only two are required to calculate the third. For simplicity and clarity we retain all three. A specific database implementation may choose to retain only two and calculate the third on the fly. ODM data loaders should only require two of the quantities to be input and should then calculate the third.

The separation of the date and time specification into two variables, LocalDateTime and UTCOffset, in the generic conceptual model may be handled differently within specific implementations. In one specific implementation these may be grouped in one text field in standard (e.g. ISO 8601) format such as YYYY-MM-DDhh:mm:ss.sss:UTCOffset (e.g. 2006-03-2516:19:56.232:-7), while in another format the date and time may be specified as the number of fractional days from an origin (e.g. Excel represents the above date as the following number 38801.6805 and allows the user to specify the format for display) with UTCOffset as a separate attribute. In general we expect specific implementations to take advantage of the representation of date time objects provided by the implementation software, but to expose the LocalDateTime and UTCOffset to users so that time may be unambiguously interpreted. In the SeriesCatalog table, begin and end times for each data series are represented by the attributes BeginDateTime, EndDateTime, BeginDateTimeUTC, and EndDateTimeUTC. The UTC offset may be derived from the difference between the UTC and local times. Because local time may change (e.g. with daylight savings) it is important during the derivation of the SeriesCatalog table that identification of the first and last records be based on UTC time and that local times be read from the corresponding records, rather than using a min or a max function on local times which can result in an error.

Support Scale
In interpreting data values that comprise a time series it is important to know the scale information associated with the data values. Blöschl and Sivapalan (1995) review the important issues. Any set of data values is quantified by a scale triplet comprising support, spacing, and extent as illustrated in Figure 6.
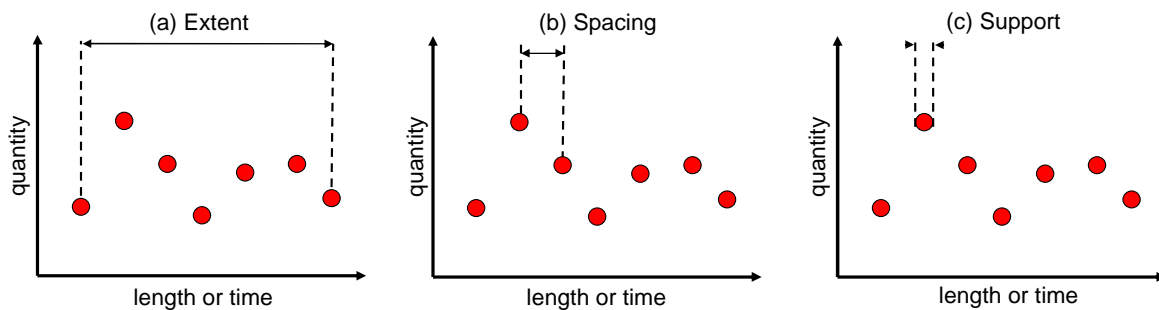
Figure 6. The scale triplet of measurements (a) extent, (b) spacing, (c) support (from Blöschl, 1996).

Extent is the full range over which the measurements occur, spacing is the spacing between measurements, and support is the averaging interval or footprint implicit in any measurement. In ODM, extent and spacing are properties of multiple measurements and are defined by the LocalDateTime or DateTimeUTC associated with data values. We have included a field called TimeSupport in the Variables table to explicitly quantify support. Figure 7 shows some of the implications associated with support, spacing, and extent in the interpretation of time series data values.

(a) spacing too large – noise (aliasing)

(b) extent too small – trend

(c) support too large – smoothing out
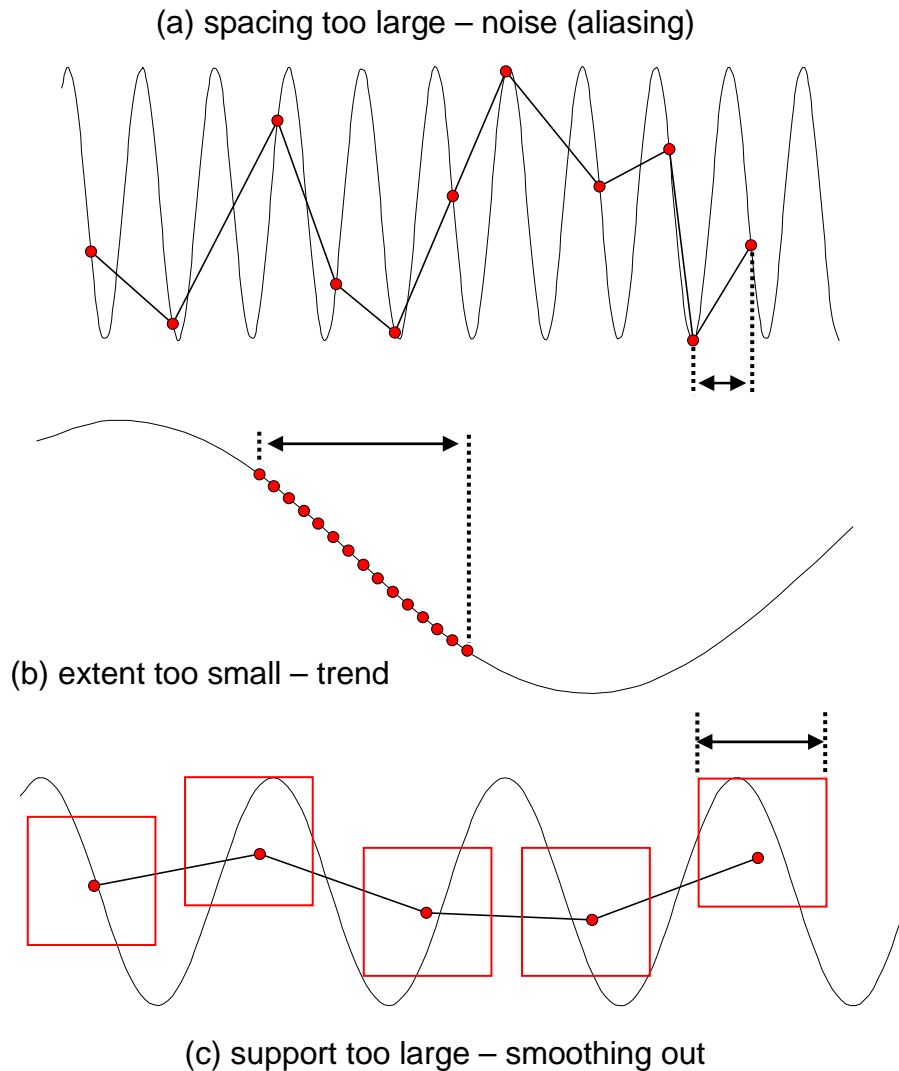
Figure 7. The effect of sampling for measurement scales not commensurate with the process scale: (a) spacing larger than the process scale causes aliasing in the data; (b) extent smaller than the process scale causes a trend in the data; (c) support larger than the process scale causes excessive smoothing in the data (adapted from Blöschl, 1996).

The concepts of scale described here apply in spatial as well as time dimensions. However, TimeSupport is only used to quantify support in the time dimension. The spatial support associated with a specific measurement method needs to be given or implied in the methods description in the Methods table. The next section indicates how time support should be specified for the different types of data.

Data Types
In the ODM, the following data types are defined. These are specified by the DataType field in the Variables table.

1. *Continuous* data – the phenomenon, such as streamflow, Q(t) is specified at a particular instant in time and measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon. Time support may be specified as 0 if the measurements are

instantaneous, or given a value that represents the time averaging inherent in the measurement method or device.

2. *Sporadic* data – the phenomenon is sampled at a particular instant in time but with a frequency that is too coarse for interpreting the record as continuous. This would be the case when the spacing is significantly larger than the support and the time scale of fluctuation of the phenomenon, such as for example infrequent water quality samples. As for continuous data, time support may be specified as 0 if the measurements are instantaneous, or given a value that represents the time averaging inherent in the measurement method or device.

3. *Cumulative* data – the data represents the cumulative value of a variable measured or calculated up to a given instant of time, such as cumulative volume of flow or cumulative precipitation:

$$V(t) = \int_0^t Q(\tau)d\tau$$, where τ represents time in the integration over the interval $[0,t]$. To

unambiguously interpret cumulative data one needs to know the time origin. In the ODM we adopt the convention of using a cumulative record with a value of zero to initialize or reset cumulative data. With this convention, cumulative data should be interpreted as the accumulation over the time interval between the date and time of the zero record and the current record at the same site position. Site position is defined by a unique combination of SiteID, VariableID, OffsetValue and OffsetType. All four of these quantities comprise the unambiguous description of the position of an observation value and there may be multiple time series associated with multiple observation positions (e.g. redundant rain gauges with different offsets) at a location. The time support for a cumulative value should be specified as 0 if the measurement of the cumulative quantity is instantaneous, or given a value that represents the time averaging inherent in the measurement of the cumulative value at the end of the period of accumulation.

4. *Incremental* data – the data value represents the incremental value of a variable over a time interval Δt such as the incremental volume of flow, or incremental precipitation:

$$\Delta V(t) = \int_t^{t+\Delta t} Q(\tau)d\tau$$. As for cumulative data, unambiguous interpretation requires knowledge of

the time increment. In the ODM we adopt the convention of using TimeSupport to specify the interval Δt, or the time interval to the next data value at the same position if TimeSupport is 0. This accommodates incremental type precipitation data that is only reported when the data value is non-zero, such as NCDC data. Such NCDC data is irregular, with the interpretation that precipitation is 0 if not reported unless qualifying comments designate otherwise. See example E.4 below for an illustration of how NCDC precipitation data is accommodated in the ODM.

5. *Average* data – the data value represents the average over a time interval, such as daily mean

discharge or daily mean temperature: $\overline{Q}(t) = \dfrac{\Delta V(t)}{\Delta t}$. The averaging interval is quantified by

TimeSupport in the case of regular data (as quantified by the IsRegular field) and by the time interval from the previous data value at the same position for irregular data.

6. *Maximum* data – the data value is the maximum value occurring at some time during a time interval, such as annual maximum discharge or a daily maximum air temperature. Again unambiguous interpretation requires knowledge of the time interval. The ODM adopts the convention that the time interval is the TimeSupport for regular data and the time interval from the previous data value at the same position for irregular data.

7. *Minimum* data – the data value is the minimum value occurring at some time during a time interval, such as 7-day low flow for a year, or the daily minimum temperature. The time interval is defined similarly to Maximum data.

8. *Constant over interval* data – the data value is a quantity that can be interpreted as constant over the time interval to the next measurement.

9. *Categorical* data – the data value is a categorical rather than continuous valued quantity. Mapping from data values to categories is through the Categories table.

We anticipate that additional data types such as median, standard deviation, variance, and others may need to be added as users work with ODM.

Beginning of Interval Reporting Time for Interval Data Values
Data types 4 to 8 above apply to data values that occur over an interval of time. The date and time reported and entered in to the ODM database associated with each interval data value is the beginning time of the observation interval. This convention was adopted to be consistent with the way dates and times are represented in most common database management systems. It should be noted that using the beginning of the interval is not consistent with the time a data logger would log an observation value. Care should be exercised in adding data to the ODM to ensure that the beginning of interval convention is followed.

Time Series Data
A considerable portion of hydrologic observations data is in the form of time series. This was why the initial model was based on the Arc Hydro Time Series Data Model. The ODM design has not specifically highlighted time series capabilities; nevertheless, the data model has inherited the key components from the Arc Hydro Time Series Data Model to give it time series capability. In particular one variable DataType is "Continuous," which is designed to indicate that the data values are collected with sufficient frequency as to be interpreted as a smooth time series. The IsRegular field also facilitates time series analysis because certain time series operations (e.g., Fourier Analysis) are predisposed to regularly sampled data. At first glance it may appear that there is redundancy between the IsRegular field and the DataType "Continuous," but we chose to keep these separate because there are regularly sampled quantities for which it is not reasonable to interpret the data values as "Continuous." For example, monthly grab samples of water quality are not continuous, but are better categorized as having DataType "Sporadic." Note that ODM does not explicitly store the time interval between measurements, nor does it indicate where a continuous series has data gaps. Both of these are required for time series analysis, but are inherently not properties of single measurements. The time interval is the time difference between sequential regular measurements, something that can be easily computed from date and time values by analysis tools. The inference of measurement gaps (and what to do about them) from date and time values we also regard as analysis functionality left for a Hydrologic Analysis System to handle.

Categorical Variables
In ODM, categorical or ordinal variables are stored in the same table as continuous valued 'real' variables through a numerical encoding of the categorical data value as a 'real' data value. The Categories table then associates, for each variable, a data value with an associated category description. This is a somewhat cumbersome construct because real valued quantities are being used as database keys. We do not see this as a significant shortcoming though, because typically, in our judgment, only a small fraction of hydrologic observations will be categorical. The Categories table stores the categories associated with categorical data values. If a Variable has a DataType that is "Categorical" then the VariableID must match one or more VariableIDs in Categories that define the mapping between DataValues and Categories. The CategoryDescription field in the Categories table defines the category.

Samples and Methods
At first glance there may appear to be redundancy between the information in the Samples table and Methods table. However, the samples table is intended to only be used where data values are derived from a physical sample that is later analyzed in a laboratory (e.g., a water chemistry sample or biological sample). The SampleID that links into the Samples table provides tracking of the specific physical sample used to derive each measurement and, by reference to information in the LabMethods table, the

laboratory methods and protocols followed. The Methods table refers to the method of field data collection, which may specify "how" a physical observation was made or collected (e.g., from an automated sampler or collected manually), but is also used to specify the measurement method associated with an in-situ measurement instrument such as a weir, turbidity sensor, dissolved oxygen sensor, humidity sensor, or temperature sensor.

Data Qualifiers

Each record in the DataValues table has an attribute called QualifierID that references the Qualifiers table. Each QualifierID in the Qualifiers table has attributes QualifierCode and QualifierDescription that provide qualifying information that can note anything unusual or problematic about individual observations such as, for example, "holding time for analysis exceeded" or "incomplete or inexact daily total." Specification of a QualifierID in the DataValues table is optional, with the inference that if a QualifierID is not specified then the corresponding data value is not qualified.

Quality Control Level Encoding

Each data value in the DataValues table has an attribute called QualityControlLevelID that references the QualityControlLevels table and is designed to record the level of quality control processing that the data value has been subjected to at the level of data series. Quality control level is one of the attributes (together with site, variable, method, and source) used to uniquely identify data series. Each quality control level is uniquely identified by its QualityControlLevelID; however, each level also has a text QualityControlLevelCode that, along with a Definition and Explanation, provides a more descriptive encoding of the quality control level. The default quality control level system used by ODM applies integer values between 0 and 4 (converted to text strings) as the QualityControlLevelCodes. Other custom systems for QualityControlLevelCodes can be used (e.g., 0.1, 0.2 to represent raw data that is progressing through a quality control work sequence, or text strings such as "Raw" or "Processed"). The following 0 – 4 QualityControlLevelCode definitions are adapted from those used by other similar systems, such as NASA, Earthscope and Ameriflux (e.g. http://ilrs.gsfc.nasa.gov/reports/ilrs_reports/9809_attach7a.html, http://public.ornl.gov/ameriflux/available.shtml accessed 3/6/2007) and are suggested so that CUAHSI ODM is consistent with the practice of other data systems:

- **QualityControlLevelCode = "0" - Raw Data**
  Raw data is defined as unprocessed data and data products that have not undergone quality control. Depending on the data type and data transmission system, raw data may be available within seconds or minutes after real-time. *Examples include real time precipitation, streamflow and water quality measurements.*

- **QualityControlLevelCode = "1" – Quality Controlled Data**
  Quality controlled data have passed quality assurance procedures such as routine estimation of timing and sensor calibration or visual inspection and removal of obvious errors. *An example is USGS published streamflow records following parsing through USGS quality control procedures.*

- **QualityControlLevelCode = "2" –Derived Products**
  Derived products require scientific and technical interpretation and include multiple-sensor data. *An example might be basin average precipitation derived from rain gages using an interpolation procedure.*

- **QualityControlLevelCode = "3" –Interpreted Products**
  These products require researcher (PI) driven analysis and interpretation, model-based interpretation using other data and/or strong prior assumptions. *An example is basin average precipitation derived from the combination of rain gages and radar return data.*

- **QualityControlLevelCode = "4" –Knowledge Products**
  These products require researcher (PI) driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions. *An example is percentages of old or new water in a hydrograph inferred from an isotope analysis.*

These definitions for quality control level are stored in the QualityControlLevels table. These definitions are recommended for use, but users can define their own quality control level system. The QualityControlLevels table is not a controlled vocabulary, but specification of a quality control level for each data value is required. Appendix B of this document provides a discussion of how to handle data versioning in terms of quality control levels (using the levels defined above), data series editing, and data series creation.

Metadata
ODM has been designed to contain all the core elements of the CUAHSI HIS metadata system (http://www.cuahsi.org/his/metadata.html) required for compliance with evolving standards such as the draft ISO 19115. In its design, the ODM embodies much record, variable, and site level metadata. Dataset and project level metadata required by these standards, such as TopicCategory, Title, and Abstract are included in a table called ISOMetaData linked to each data source.

Reference Documents
The Methods, Sources, LabMethods and ISOMetaData tables contain fields that can be used to store links to source or reference information. At the general conceptual level of the ODM we do not specify how, or in what form these links to references or sources should be implemented. Options include using URLs or storing entire documents in the database. If external URLs are used it will be important as the database grows and is used over time to ensure that links or URLs included are stable. An alternative approach to external links is to exploit the capability of modern databases to store entire digital documents, such as an html or xml page, PDF document, or raw data file, within a field in the database. The capability therefore exists to instead have these links refer to a separate table that would actually contain this metadata information, instead of housing it in a separate digital library. There is some merit in this because then any data exported in ODM format could take with it the associated metadata required to completely define it as well as the raw data upon which it is derived. However, this has the disadvantage of increasing (perhaps substantially) the size of database file containing the data and being distributed to users.

Controlled Vocabularies
The following tables in the ODM are tables where controlled vocabularies for the fields are required to maintain consistency and avoid the use of synonyms that can lead to ambiguity:

- CensorCodeCV
- DataTypeCV
- GeneralCategoryCV
- SampleMediumCV
- SampleTypeCV
- SpatialReferences
- SpeciationCV
- TopicCategoryCV
- Units
- ValueTypeCV
- VariableNameCV

- VerticalDatumCV

The initial contents of these controlled vocabularies are specified in the Microsoft SQL Server 2005 blank schema for the ODM. However, the ODM controlled vocabularies are dynamic. A central repository of current ODM controlled vocabulary terms is maintained on the ODM Website at http://water.usu.edu/cuahsi/odm/, together with the most recent version of the ODM SQL Server 2005 blank schema, this design specifications document, and other tools for working with ODM. Users can submit new terms for the controlled vocabularies and can request changes to existing terms using functionality available on the ODM website (http://water.usu.edu/cuahsi/odm/). Functionality for updating local controlled vocabulary tables with new terms from the central ODM controlled vocabulary repository is provided in the ODM Tools software application, which is also available from the ODM website. The CUAHSI HIS team welcomes input on the controlled vocabularies.

Examples

The following examples show the capability of ODM to store different types of point observations. It is not possible in examples such as these to present all of the field values for all the tables. Because of this, the examples present selected fields and tables chosen to illustrate key capabilities of the data model. Refer to Appendix A for the complete definition of table and field contents.

Streamflow - Gage Height and Discharge

Figure E.1 illustrates how both stream gage height measurements and the associated discharge estimates derived from the gage height measurements can be stored in the ODM. Note that gage height in feet and discharge in cubic feet per second are both in the same data table but with different VariableIDs that reference the Variables table, which specifies the VariableName, Units, and other quantities associated with these data values. The link between VariableID in the DataValues table and Variables table is shown. In this example, discharge measurements are derived from gage height (stage) measurements through a rating curve. The MethodID associated with each discharge record references into the Methods table that describes this and provides a URL that should contain metadata details for this method. The DerivedFromID in the DataValues table references into the DerivedFrom table that references back to the corresponding gage height in the DataValues table from which the discharge was derived.
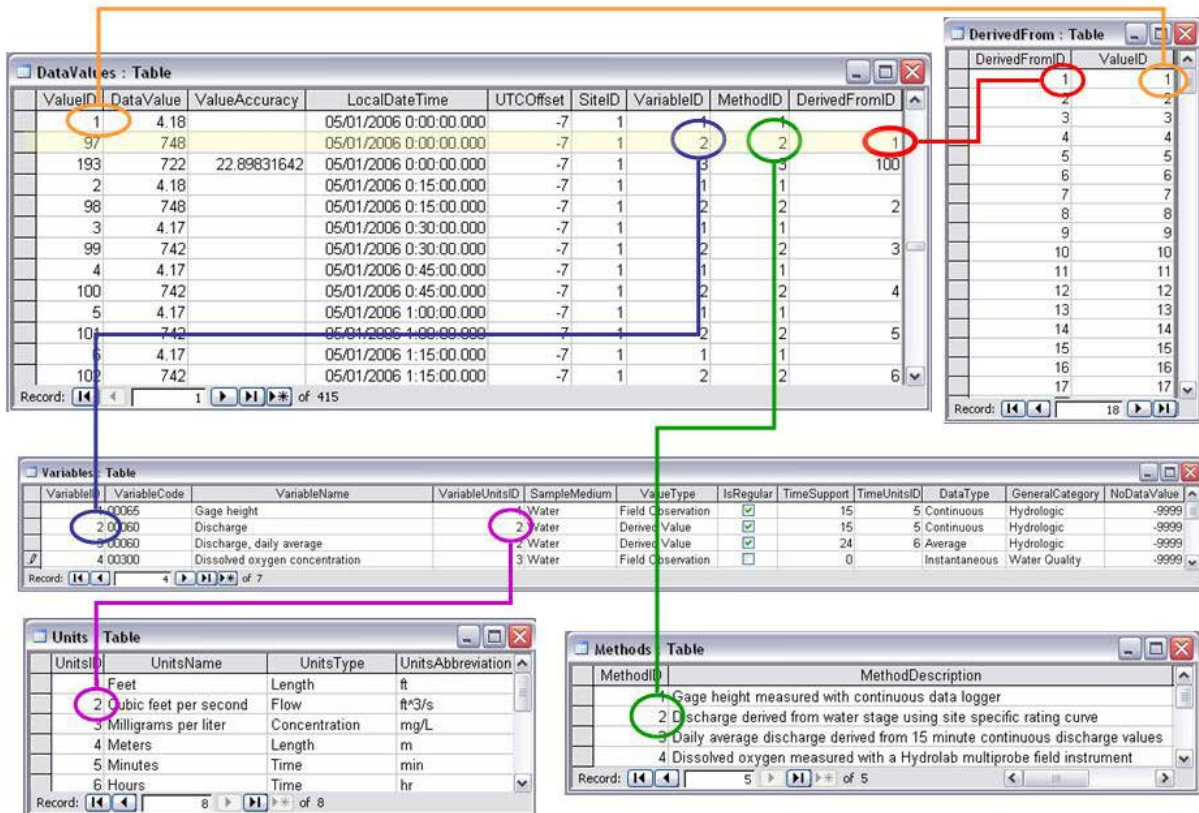
Figure E.1. Excerpts from tables illustrating the population of ODM with streamflow gage height (stage) and discharge data.

Streamflow - Daily Average Discharge

Daily average streamflow is reported as an average of continuous 15 minute interval data values. Figure E.2 shows excerpts from tables illustrating the population of ODM with both the continuous discharge values and derived daily averages. The record giving the single daily average discharge with a value of 722 ft^3/s in the DataValues table has a DerivedFromID of 100. This refers to multiple records in the DerivedFrom table, with associated ValueIDs 97, 98, 99, … 113 shown. These refer to the specific 15 minute discharge values in the DataValues table used to derive the average daily discharge. VariableID in the DataValues table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of ft^3/s from UnitsID referencing in to the Units table. MethodID in the DataValues table identifies the appropriate record in the Methods table specifying that the method used to obtain this data value was daily averaging.
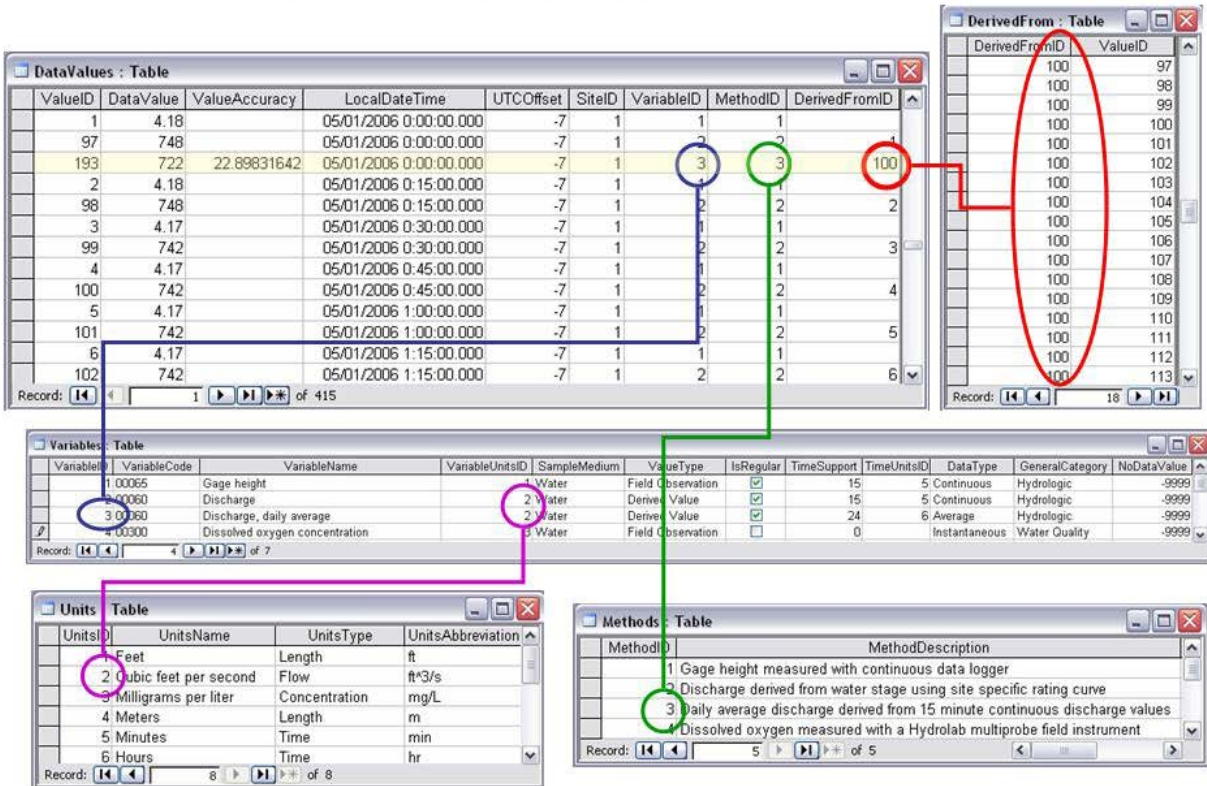
Figure E.2. Excerpts from tables illustrating the population of ODM with daily average discharge derived from 15 minute discharge values.

Water Chemistry from a Profile in a Lake

Reservoir profile measurements provide an example of the logical grouping of data values and data values that have an offset in relationship to the location of the monitoring site. These measurements may be made simultaneously (by multiple instruments in the water column) or over a short time period (one instrument that is lowered from top to bottom). Figure E.3 shows an example of how these data would be stored in ODM. The OffsetTypes table and OffsetValue attribute is used to quantify the depth offset associated with each measurement. Each of the data values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType the offset is "Depth below water surface." The OffsetTypes table references into the Units table indicating that the OffsetUnits are meters, so OffsetValue in the DataValues table is in units of meters depth below the water surface. Each of the data values shown also has a VariableID that in the Variables table indicates that the variable measured was dissolved oxygen concentration in units of mg/L. Each of the data values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The data values shown are part of a logical group of data values representing the water chemistry profile in a lake. This is represented using the Groups table and GroupDescriptions table. The Groups table associates GroupID 1 with each of the ValueIDs of the data values belonging to the group. A description of this group is given in the GroupDescriptions table.
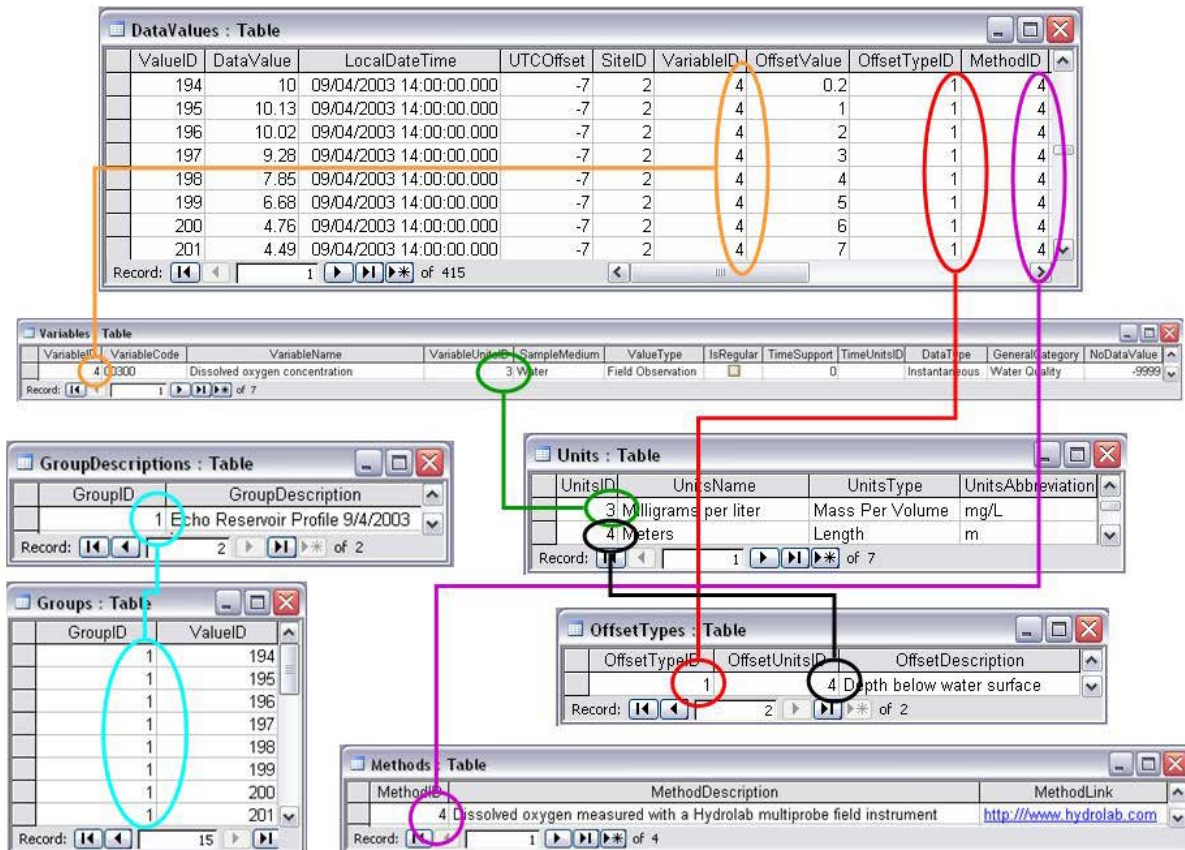
Figure E.3. Excerpts from tables illustrating the population of ODM with water chemistry data.

NCDC Precipitation Data

Figure E.4 illustrates the representation of NCDC 15 minute precipitation data by ODM. The data includes 15 minute incremental data values as well as daily totals. Separate records in the Variables table are used for the 15 minute or daily total values. These data are reported at irregular intervals and only logged for time periods for which precipitation is non zero. This is accommodated by setting the IsRegular attribute associated with the variable to "False" and specifying the TimeSupport value as 15 or 24 and the TimeUnits as "Minutes" or "Hours". The DataType of "Incremental" is used to indicate that these are incremental data values defined over the TimeSupport interval. The convention for incremental data (see above) is that when the time support is specified, it specifies the increment for irregular incremental data. When time support is specified as 0 it means the increment is from the previous data value at the same site position. Data qualifiers indicate periods where the data is missing. The method associated with each precipitation variable documents the convention that zero precipitation periods are not logged in this data acquired from NCDC. A data qualifier is also used to flag days where the precipitation total is incomplete due to the record being missing during part of the day.

B-23
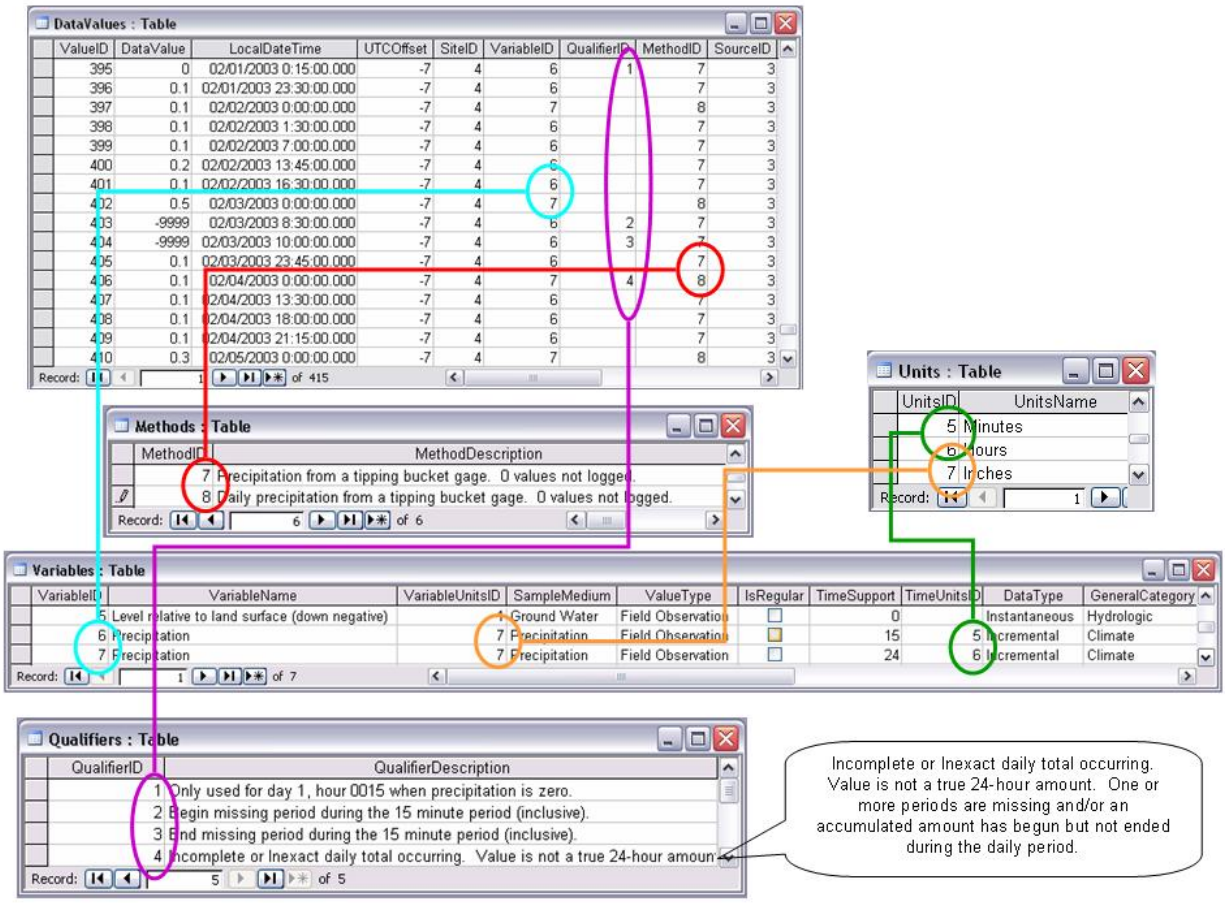
Figure E.4. Excerpts from tables illustrating the population of the ODM with NCDC Precipitation Data.

Groundwater Level Data

The following is an example of how groundwater level data can be stored in ODM. In this example, the data values are the water table level relative to the ground surface reported as negative values. This example shows multiple data values of a single variable at a single site made by a single source that have been quality controlled as indicated by the QualityControlLevelID field in the QualityControlLevels table. The SiteID field in the DataValues table indicates the site in the Sites table that gives the location information about the monitoring site. In this case, the elevation is with respect to the NGVD29 datum as indicated in the VerticalDatum field, and latitude and longitude are with respect to the NAD27 datum as indicated in the SpatialReferences table. The VariableID field in the DataValues table references the appropriate record in the Variables table indicating information about the variable. The SourceID field in the DataValues table references the appropriate record in the Sources table giving information about the source of the data.
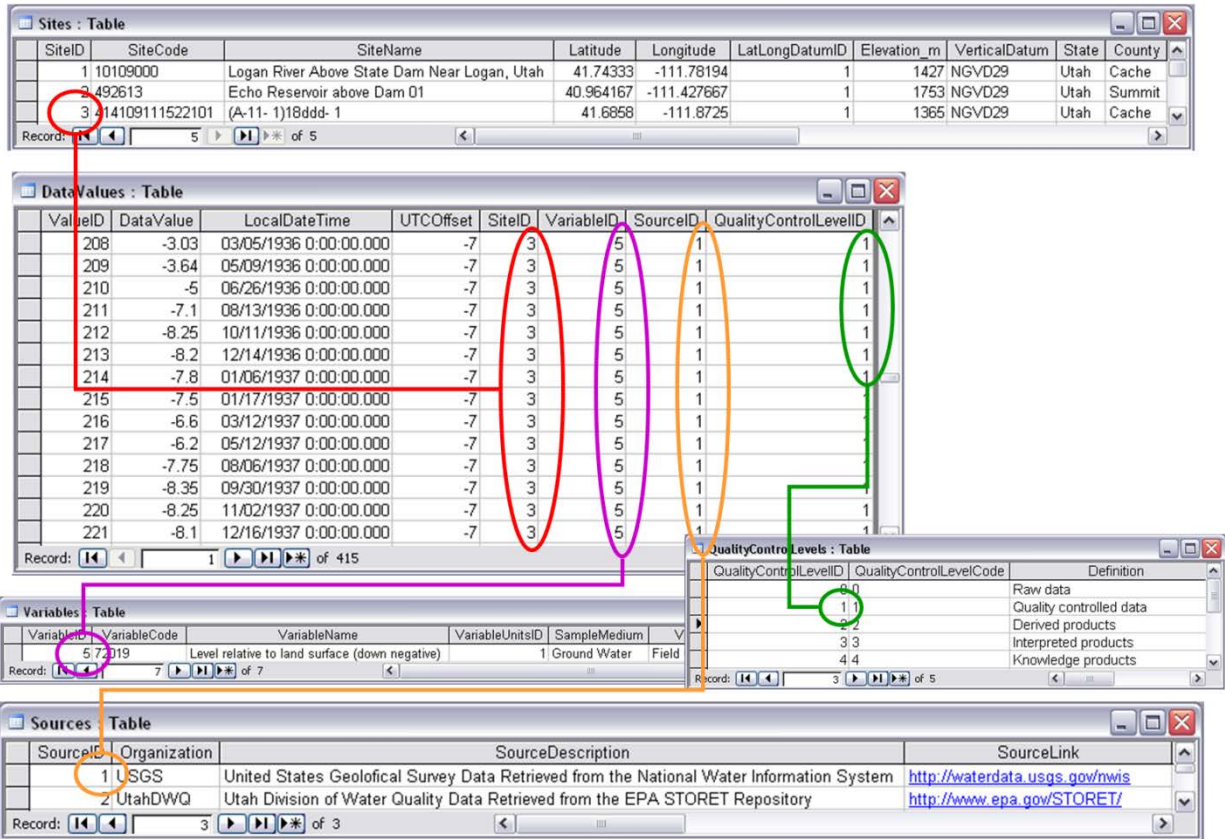
Figure E.5. Excerpts from tables illustrating the population of the ODM with irregularly sampled groundwater level data.

References

AIAA, (1995), Assessment of Wind Tunnel Data Uncertainty, American Institute of Aeronautics and Astronautics: AIAA S-071-1995.

Blöschl, G., (1996), Scale and Scaling in Hydrology, Habilitationsschrift, Weiner Mitteilungen Wasser Abwasser Gewasser, Wien, 346 p.

Blöschl, G. and M. Sivapalan, (1995), "Scale Issues in Hydrological Modelling: A Review," Hydrological Processes, 9(1995): 251-290.

Horsburgh, J. S., D. G. Tarboton and D. R. Maidment, (2005), "A Community Data Model for Hydrologic Observations, Chapter 6," in Hydrologic Information System Status Report, Version 1, Edited by D. R. Maidment, p.102-135, http://www.cuahsi.org/his/docs/HISStatusSept15.pdf.

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky, (2008), A relational model for environmental and water resources data, Water Resources Research, Vol. 44, W05406, doi:10.1029/2007WR006392.

Maidment, D. R., ed. (2002), Arc Hydro GIS for Water Resources, ESRI Press, Redlands, CA, 203 p.

Maidment, D. R., (2005), "A Data Model for Hydrologic Observations." Paper prepared for presentation at the CUAHSI Hydrologic Information Systems Symposium, University of Texas at Austin. March 7, 2005.

Tarboton, D. G., (2005), "Review of Proposed CUAHSI Hydrologic Information System Hydrologic Observations Data Model." Utah State University. May 5, 2005.

Appendix A. Observations Data Model Table and Field Structure

The following is a description of the tables in the observations data model, a listing of the fields contained in each table, a description of the data contained in each field and its data type, examples of the information to be stored in each field where appropriate, specific constraints imposed on each field, and discussion on how each field should be populated. Values in the example column should not be considered to be inclusive of all potential values, especially in the case of fields that require a controlled vocabulary. We anticipate that these controlled vocabularies will need to be extended and adjusted. Tables appear in alphabetical order.

Each table below includes a "Constraint" column. The value in this column designates each field in the table as one of the following:

Mandatory (M) – A value in this field is mandatory and cannot be NULL.
Optional (O) – A value in this field is optional and can be NULL.
Programmatically derived (P) – Inherits from the source field. The value in this field should be automatically populated as the result of a query and is not required to be input by the user.

Additional constraints are documented where appropriate in the Constraint column. In addition, where appropriate, each table contains a "Default Value" column. The value in this column is the default value for the associated field. The default value specifies the convention that should be followed when a value for the field is not specified. Below each table is a discussion of the rules and best practices that should be used in populating each table within ODM.

Table: Categories

The Categories table defines the categories for categorical variables. Records are required for variables where DataType is specified as "Categorical." Multiple entries for each VariableID, with different DataValues provide the mapping from DataValue to category description.

Field Name	DataType	Description	Examples	Constraint
VariableID	Integer	Integer identifier that references the Variables record of a categorical variable.	45	M Foreign key
DataValue	Real	Numeric value that defines the category	1.0	M
CategoryDescription	Text (Unlimited)	Definition of categorical variable value	"Cloudy"	M

The following rules and best practices should be used in populating this table:

1. Although all of the fields in this table are mandatory, they need only be populated if categorical data are entered into the database. If there are no categorical data in the DataValues table, this table will be empty.
2. This table should be populated before categorical data values are added to the DataValues table.

Table: CensorCodeCV

The CensorCodeCV table contains the controlled vocabulary for censor codes. Only values from the Term field in this table can be used to populate the CensorCode field of the DataValues table.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for CensorCode.	"lt", "gt", "nc"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of CensorCode controlled vocabulary term. The definition is optional if the term is self explanatory.	"less than", "greater than", "not censored"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: DataTypeCV

The DataTypeCV table contains the controlled vocabulary for data types. Only values from the Term field in this table can be used to populate the DataType field in the Variables table.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for DataType.	"Continuous"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of DataType controlled vocabulary term. The definition is optional if the term is self explanatory.	"A quantity specified at a particular instant in time measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon."	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: DataValues

The DataValues table contains the actual data values.

Field Name	Data Type	Description	Example	Constraint	Default Value
ValueID	Integer Identity	Unique integer identifier for each data value.	43	M Unique Primary key	
DataValue	Real	The numeric value of the observation. For Categorical variables, a number is stored here. The Variables table has DataType as Categorical and the Categories table maps from the DataValue onto Category Description.	34.5	M	
ValueAccuracy	Real	Numeric value that describes the measurement accuracy of the data value. If not given, it is interpreted as unknown.	4	O	NULL
LocalDateTime	Date/Time	Local date and time at which the data value was observed. Represented in an implementation specific format.	9/4/2003 7:00:00 AM	M	
UTCOffset	Real	Offset in hours from UTC time of the corresponding LocalDateTime value.	-7	M	
DateTimeUTC	Date/Time	Universal UTC date and time at which the data value was observed. Represented in an implementation specific format.	9/4/2003 2:00:00 PM	M	
SiteID	Integer	Integer identifier that references the site at which the observation was measured. This links data values to their locations in the Sites table.	3	M Foreign key	
VariableID	Integer	Integer identifier that references the variable that was measured. This links data values to their variable in the Variables table.	5	M Foreign key	
OffsetValue	Real	Distance from a datum or control point to the point at which a data value was observed. If not given the OffsetValue is inferred to be 0, or not relevant/necessary.	2.1	O	NULL = No Offset
OffsetTypeID	Integer	Integer identifier that references the measurement offset type in the OffsetTypes table.	3	O Foreign key	NULL = No Offset
CensorCode	Text (50)	Text indication of whether the data value is censored from the CensorCodeCV controlled vocabulary.	"nc"	M Foreign key	"nc" = Not Censored
QualifierID	Integer	Integer identifier that references the Qualifiers table. If Null, the data value is inferred to not be qualified.	4	O Foreign key	NULL
MethodID	Integer	Integer identifier that references method used to generate the data value in the Methods table.	3	M Foreign key	0 = No method specified

Field Name	Data Type	Description	Example	Constraint	Default Value
SourceID	Integer	Integer identifier that references the record in the Sources table giving the source of the data value.	5	M Foreign key	
SampleID	Integer	Integer identifier that references into the Samples table. This is required only if the data value resulted from a physical sample processed in a lab.	7	O Foreign key	NULL
DerivedFromID	Integer	Integer identifier for the derived from group of data values that the current data value is derived from. This refers to a group of derived from records in the DerivedFrom table. If NULL, the data value is inferred to not be derived from another data value.	5	O	NULL
QualityControlLevelID	Integer	Integer identifier giving the level of quality control that the value has been subjected to. This references the QualityControlLevels table.	1	M Foreign key	-9999 = Unknown

The following rules and best practices should be used in populating this table:

1. ValueID is the primary key, is mandatory, and cannot be NULL. This field should be implemented as an autonumber/identity field. When data values are added to this table, a unique integer ValueID should be assigned to each data value by the database software such that the primary key constraint is not violated.
2. Each record in this table must be unique. This is enforced by a unique constraint across all of the fields in this table (excluding ValueID) so that duplicate records are avoided.
3. The LocalDateTime, UTCOffset, and DateTimeUTC must all be populated. Care must be taken to ensure that the correct UTCOffset is used, especially in areas that observe daylight saving time. If LocalDateTime and DateTimeUTC are given, the UTCOffset can be calculated as the difference between the two dates. If LocalDateTime and UTCOffset are given, DateTimeUTC can be calculated.
4. SiteID must correspond to a valid SiteID from the Sites table. When adding data for a new site to the ODM, the Sites table should be populated prior to adding data values to the DataValues table.
5. VariableID must correspond to a valid VariableID from the Variables table. When adding data for a new variable to the ODM, the Variables table should be populated prior to adding data values for the new variable to the DataValues table.
6. OffsetValue and OffsetTypeID are optional because not all data values have an offset. Where no offset is used, both of these fields should be set to NULL indicating that the data values do not have an offset. Where an OffsetValue is specified, an OffsetTypeID must also be specified and it must refer to a valid OffsetTypeID in the OffsetTypes table. The OffsetTypes table should be populated prior to adding data values with a particular OffsetTypeID to the DataValues table.
7. CensorCode is mandatory and cannot be NULL. A default value of "nc" is used for this field. Only Terms from the CensorCodeCV table should be used to populate this field.
8. The QualifierID field is optional because not all data values have qualifiers. Where no qualifier applies, this field should be set to NULL. When a QualifierID is specified in this field it must refer to a valid QualifierID in the Qualifiers table. The Qualifiers table should be populated prior to adding data values with a particular QualifierID to the DataValues Table.

9. MethodID must correspond to a valid MethodID from the Methods table and cannot be NULL. A default value of 0 is used in the case where no method is specified or the method used to create the observation is unknown. The Methods table should be populated prior to adding data values with a particular MethodID to the DataValues table.

10. SourceID must correspond to a valid SourceID from the Sources table and cannot be NULL. The Sources table should be populated prior to adding data values with a particular SourceID to the DataValues table.

11. SampleID is optional and should only be populated if the data value was generated from a physical sample that was sent to a laboratory for analysis. The SampleID must correspond to a valid SampleID in the Samples table, and the Samples table should be populated prior to adding data values with a particular SampleID to the DataValues table.

12. DerivedFromID is optional and should only be populated if the data value was derived from other data values that are also stored in the ODM database.

13. QualityControlLevelID is mandatory, cannot be NULL, and must correspond to a valid QualityControlLevelID in the QualityControlLevels table. A default value of -9999 is used for this field in the event that the QualityControlLevelID is unknown. The QualityControlLevels table should be populated prior to adding data values with a particular QualityControlLevelID to the DataValues table.

Table: DerivedFrom

The DerivedFrom table contains the linkage between derived data values and the data values that they were derived from.

Field Name	Data Type	Description	Examples	Constraint
DerivedFromID	Integer	Integer identifying the group of data values from which a quantity is derived.	3	M
ValueID	Integer	Integer identifier referencing data values that comprise a group from which a quantity is derived. This corresponds to ValueID in the DataValues table.	1,2,3,4,5	M

The following rules and best practices should be used in populating this table:

1. Although all of the fields in this table are mandatory, they need only be populated if derived data values and the data values that they were derived from are entered into the database. If there are no derived data in the DataValues table, this table will be empty.

Table: GeneralCategoryCV

The GeneralCategoryCV table contains the controlled vocabulary for the general categories associated with Variables. The GeneralCategory field in the Variables table can only be populated with values from the Term field of this controlled vocabulary table.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for GeneralCategory.	"Hydrology"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of GeneralCategory controlled vocabulary term. The definition is optional if the term is self explanatory.	"Data associated with hydrologic variables or processes."	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: GroupDescriptions

The GroupDescriptions table lists the descriptions for each of the groups of data values that have been formed.

Field Name	Data Type	Description	Example	Constraint
GroupID	Integer Identity	Unique integer identifier for each group of data values that has been formed. This also references to GroupID in the Groups table.	4	M Unique Primary key
GroupDescription	Text (unlimited)	Text description of the group.	"Echo Reservoir Profile 7/7/2005"	O

The following rules and best practices should be used in populating this table:

1. This table will only be populated if groups of data values have been created in the ODM database.
2. The GroupID field is the primary key, must be a unique integer, and cannot be NULL. It should be implemented as an auto number/identity field.
3. The GroupDescription can be any text string that describes the group of observations.

Table: Groups

The Groups table lists the groups of data values that have been created and the data values that are within each group.

Field Name	Data Type	Description	Example	Constraint
GroupID	Integer	Integer ID for each group of data values that has been formed.	4	M Foreign key
ValueID	Integer	Integer identifier for each data value that belongs to a group. This corresponds to ValueID in the DataValues table	2,3,4	M Foreign key

The following rules and best practices should be used in populating this table:

1. This table will only be populated if groups of data values have been created in the ODM database.
2. The GroupID field must reference a valid GroupID from the GroupDescriptions table, and the GroupDescriptions table should be populated for a group prior to populating the Groups table.

Table: ISOMetadata

The ISOMetadata table contains dataset and project level metadata required by the CUAHSI HIS metadata system (http://www.cuahsi.org/his/documentation.html) for compliance with standards such as the draft ISO 19115 or ISO 8601. The mandatory fields in this table must be populated to provide a complete set of ISO compliant metadata in the database.

Field Name	Data Type	Description	Example	Constraint	Default Value
MetadataID	Integer Identity	Unique integer ID for each metadata record.	4	M Unique Primary key	
TopicCategory	Text (255)	Topic category keyword that gives the broad ISO19115 metadata topic category for data from this source. The controlled vocabulary of topic category keywords is given in the TopicCategoryCV table.	"inlandWaters"	M Foreign key	"Unknown"
Title	Text (255)	Title of data from a specific data source.		M Cannot contain tab, line feed, or carriage return characters	"Unknown"
Abstract	Text (unlimited)	Abstract of data from a specific data source.		M	"Unknown"
ProfileVersion	Text (255)	Name of metadata profile used by the data source	"ISO8601"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
MetadataLink	Text (500)	Link to additional metadata reference material.		O	NULL

The following rules and best practices should be used in populating this table:

1. The MetadataID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
2. All of the fields in this table are mandatory and cannot be NULL except for the MetadataLink field.
3. The TopicCategory field should only be populated with terms from the TopicCategoryCV table. The default controlled vocabulary term is "Unknown."

4. The Title field should be populated with a brief text description of what the referenced data represent. This field can be populated with "Unknown" if there is no title for the data.
5. The Abstract field should be populated with a more complete text description of the data that the metadata record references. This field can be populated with "Unknown" if there is no abstract for the data.
6. The ProfileVersion field should be populated with the version of the ISO metadata profile that is being used. This field can be populated with "Unknown" if there is no profile version for the data.
7. One record with a MetadataID = 0 should exist in this table with TopicCategory, Title, Abstract, and ProfileVersion = "Unknown" and MetadataLink = NULL. This record should be the default value for sources with unknown/unspecified metadata.

Table: LabMethods

The LabMethods table contains descriptions of the laboratory methods used to analyze physical samples for specific constituents.

Field Name	Data Type	Description	Example	Constraint	Default Value
LabMethodID	Integer Identity	Unique integer identifier for each laboratory method. This is the key used by the Samples table to reference a laboratory method.	6	M Unique Primary key	
LabName	Text (255)	Name of the laboratory responsible for processing the sample.	"USGS Atlanta Field Office"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
LabOrganization	Text (255)	Organization responsible for sample analysis.	"USGS"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
LabMethodName	Text (255)	Name of the method and protocols used for sample analysis.	"USEPA-365.1"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
LabMethodDescription	Text (unlimited)	Description of the method and protocols used for sample analysis.	"Processed through Model *** Mass Spectrometer"	M	"Unknown"
LabMethodLink	Text (500)	Link to additional reference material on the analysis method.		O	NULL

The following rules and best practices should be used when populating this table:

1. The LabMethodID field is the primary key, must be a unique integer, and cannot be NULL. It should be implemented as an auto number/identity field.
2. All of the fields in this table are required and cannot be null except for the LabMethodLink.
3. The default value for all of the required fields except for the LabMethodID is "Unknown."
4. A single record should exist in this table where the LabMethodID = 0 and the LabName, LabOrganization, LabMethdodName, and LabMethodDescription fields are equal to "Unknown" and the LabMethodLink = NULL. This record should be used to identify samples in the Samples table for which nothing is known about the laboratory method used to analyze the sample.

Table: Methods

The Methods table lists the methods used to collect the data and any additional information about the method.

Field Name	Data Type	Description	Example	Constraint	Default Value
MethodID	Integer Identity	Unique integer ID for each method.	5	M Unique Primary key	
MethodDescription	Text (unlimited)	Text description of each method.	"Specific conductance measured using a Hydrolab" or "Streamflow measured using a V notch weir with dimensions xxx"	M	
MethodLink	Text (500)	Link to additional reference material on the method.		O	NULL

The following rules and best practices should be used when populating this table:

1. The MethodID field is the primary key, must be a unique integer, and cannot be NULL.
2. There is no default value for the MethodDescription field in this table. Rather, this table should contain a record with MethodID = 0, MethodDescription = "Unknown", and MethodLink = NULL. A MethodID of 0 should be used as the MethodID for any data values for which the method used to create the value is unknown (i.e., the default value for the MethodID field in the DataValues table is 0).
3. Methods should describe the manner in which the observation was collected (i.e., collected manually, or collected using an automated sampler) or measured (i.e., measured using a temperature sensor or measured using a turbidity sensor). Details about the specific sensor models and manufacturers can be included in the MethodDescription.

Table: ODM Version

The ODM Version table has a single record that records the version of the ODM database. This table must contain a valid ODM version number. This table will be pre-populated and should not be edited.

Field Name	Data Type	Description	Example	Constraint
VersionNumber	Text (50)	String that lists the version of the ODM database.	"1.1"	M Cannot contain tab, line feed, or carriage return characters

Table: OffsetTypes

The OffsetTypes table lists full descriptive information for each of the measurement offsets.

Field Name	Data Type	Description	Example	Constraint
OffsetTypeID	Integer Identity	Unique integer identifier that identifies the type of measurement offset.	2	M Unique Primary key
OffsetUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the OffsetValue.	1	M Foreign key
OffsetDescription	Text (unlimited)	Full text description of the offset type.	"Below water surface" "Above Ground Level"	M

The following rules and best practices should be followed when populating this table:

1. Although all three fields in this table are mandatory, this table will only be populated if data values measured at an offset have been entered into the ODM database.
2. The OffsetTypeID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
3. The OffsetUnitsID field should reference a valid ID from the UnitsID field in the Units table. Because the Units table is a controlled vocabulary, only units that already exist in the Units table can be used as the units of the offset.
4. The OffsetDescription field should be filled in with a complete text description of the offset that provides enough information to interpret the type of offset being used. For example, "Distance from stream bank" is ambiguous because it is not known which bank is being referred to.

Table: Qualifiers

The Qualifiers table contains data qualifying comments that accompany the data.

Field Name	Data Type	Description	Example	Constraint	Default Value
QualifierID	Integer Identity	Unique integer identifying the data qualifier.	3	M Unique Primary key	

QualifierCode	Text (50)	Text code used by organization that collects the data.	"e" (for estimated) or "a" (for approved) or "p" (for provisional)	O Cannot contain space, tab, line feed, or carriage return characters	NULL
QualifierDescription	Text (unlimited)	Text of the data qualifying comment.	"Holding time for sample analysis exceeded"	M	

This table will only be populated if data values that have data qualifying comments have been added to the ODM database. The following rules and best practices should be used when populating this table:

1. The QualifierID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.

Table: QualityControlLevels

The QualityControlLevels table contains the quality control levels that are used for versioning data within the database.

Field Name	Data Type	Description	Example	Constraint
QualityControlLevelID	Integer Identity	Unique integer identifying the quality control level.	0, 1, 2, 3, 4, 5	M Unique Primary key
QualityControlLevelCode	Text (50)	Code used to identify the level of quality control to which data values have been subjected.	"1", "1.1", "Raw", "QC Checked"	M Cannot contain tab, line feed, or carriage return characters
Definition	Text (255)	Definition of Quality Control Level.	"Raw Data", "Quality Controlled Data"	M Cannot contain tab, line feed, or carriage return characters
Explanation	Text (unlimited)	Explanation of Quality Control Level	"Raw data is defined as unprocessed data and data products that have not undergone quality control."	M

This table is pre-populated with quality control levels 0 through 4 within the ODM. The following rules and best practices should be used when populating this table:

1. The QualityControlLevelID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
2. It is suggested that the pre-populated system of quality control level codes (i.e., QualityControlLevelCodes 0 – 4) be used. If the pre-populated list is not sufficient, new quality control levels can be defined. A quality control level code of -9999 is suggested for data whose quality control level is unknown.

Table: SampleMediumCV

The SampleMediumCV table contains the controlled vocabulary for sample media.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for sample media.	"Surface Water"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of sample media controlled vocabulary term. The definition is optional if the term is self explanatory.	"Sample taken from surface water such as a stream, river, lake, pond, reservoir, ocean, etc."	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: Samples

The Samples table gives information about physical samples analyzed in a laboratory.

Field Name	Data Type	Description	Example	Constraint	Default Value
SampleID	Integer Identity	Unique integer identifier that identifies each physical sample.	3	M Unique Primary key	
SampleType	Text (255)	Controlled vocabulary specifying the sample type from the SampleTypeCV table.	"FD", "PB", "SW", "Grab Sample"	M Foreign key	"Unknown"
LabSampleCode	Text (50)	Code or label used to identify and track lab sample or sample container (e.g. bottle) during lab analysis.	"AB-123"	M Unique Cannot contain tab, line feed, or carriage return characters	

Field Name	Data Type	Description	Example	Constraint	Default Value
LabMethodID	Integer	Unique identifier for the laboratory method used to process the sample. This references the LabMethods table.	4	M Foreign key	0 = Nothing known about lab method

The following rules and best practices should be followed when populating this table:

1. This table will only be populated if data values associated with physical samples are added to the ODM database.
2. The SamplID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
3. The SampleType field should be populated using terms from the SampleTypeCV table. Where the sample type is unknown, a default value of "Unknown" can be used.
4. The LabSampleCode should be a unique text code used by the laboratory to identify the sample. This field is an alternate key for this table and should be unique.
5. The LabMethodID must reference a valid LabMethodID from the LabMethods table. The LabMethods table should be populated with the appropriate laboratory method information prior to adding records to this table that reference that laboratory method. A default value of 0 for this field indicates that nothing is known about the laboratory method used to analyze the sample.

Table: SampleTypeCV

The SampleTypeCV table contains the controlled vocabulary for sample type.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for sample type.	"FD", "PB", "Grab Sample"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of sample type controlled vocabulary term. The definition is optional if the term is self explanatory.	"Foliage Digestion", "Precipitation Bulk"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: SeriesCatalog

The SeriesCatalog table lists each separate data series in the database for the purposes of identifying or displaying what data are available at each site and to speed simple queries without querying the main DataValues table. Unique site/variable combinations are defined by unique combinations of SiteID, VariableID, MethodID, SourceID, and QualityControlLevelID.

This entire table should be programmatically derived and should be updated every time data is added to the database. Constraints on each field in the SeriesCatalog table are dependent upon the constraints on the fields in the table from which those fields originated.

Field Name	Data Type	Description	Example	Constraint
SeriesID	Integer Identity	Unique integer identifier for each data series.	5	P Primary key
SiteID	Integer	Site identifier from the Sites table.	7	P
SiteCode	Text (50)	Site code used by organization that collects the data.	"1002000"	P
SiteName	Text (255)	Full text name of sampling site.	"Logan River"	P
VariableID	Integer	Integer identifier for each Variable that references the Variables table.	4	P
VariableCode	Text (50)	Variable code used by the organization that collects the data.	"00060"	P
VariableName	Text (255)	Name of the variable from the variables table.	"Temperature"	P
Speciation	Text (255)	Code used to identify how the data value is expressed (i.e., total phosphorus concentration expressed *as P*). This should be from the SpeciationCV controlled vocabulary table.	"P", "N", "NO3"	P
VariableUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the data value.	5	P
VariableUnitsName	Text (255)	Full text name of the variable units from the UnitsName field in the Units table.	"milligrams per liter"	P
SampleMedium	Text (255)	The medium of the sample. This should be from the SampleMediumCV controlled vocabulary table.	"Surface Water"	P
ValueType	Text (255)	Text value indicating what type of data value is being recorded. This should be from the ValueTypeCV controlled vocabulary table.	"Field Observation"	P
TimeSupport	Real	Numerical value that indicates the time support (or temporal footprint) of the data values. 0 is used to indicate data values that are instantaneous. Other values indicate the time over which the data values are implicitly or explicitly averaged or aggregated.	0, 24	P

Field Name	Data Type	Description	Example	Constraint
TimeUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the time support. If TimeSupport is 0, indicating an instantaneous observation, a unit needs to still be given for completeness, although it is somewhat arbitrary.	4	P
TimeUnitsName	Text (255)	Full text name of the time support units from the UnitsName field in the Units table.	"hours"	P
DataType	Text (255)	Text value that identifies the data as one of several types from the DataTypeCV controlled vocabulary table.	"Continuous" "Instantaneous" "Cumulative" "Incremental" "Average" "Minimum" "Maximum" "Constant Over Interval" "Categorical"	P
GeneralCategory	Text (255)	General category of the variable from the GeneralCategoryCV table.	"Water Quality"	P
MethodID	Integer	Integer identifier that identifies the method used to generate the data values and references the Methods table.	2	P
MethodDescription	Text (unlimited)	Full text description of the method used to generate the data values.	"Specific conductance measured using a Hydrolab" or "Streamflow measured using a V notch weir with dimensions xxx"	P
SourceID	Integer	Integer identifier that identifies the source of the data values and references the Sources table.	5	P
Organization	Text (255)	Text description of the source organization from the Sources table.	"USGS"	P
SourceDescription	Text (unlimited)	Text description of the data source from the Sources table.	"Text file retrieved from the EPA STORET system indicating data originally from Utah Division of Water Quality"	P

Field Name	Data Type	Description	Example	Constraint
Citation	Text (unlimited)	Text string that give the citation to be used when the data from each source are referenced.	"Slaughter, C. W., D. Marks, G. N. Flerchinger, S. S. Van Vactor and M. Burgess, (2001), "Thirty-five years of research data collection at the Reynolds Creek Experimental Watershed, Idaho, United States," Water Resources Research, 37(11): 2819-2823."	P
QualityControlLevelID	Integer	Integer identifier that indicates the level of quality control that the data values have been subjected to.	0,1,2,3,4	P
QualityControlLevelCode	Text (50)	Code used to identify the level of quality control to which data values have been subjected.	"1", "1.1", "Raw", "QC Checked"	P
BeginDateTime	Date/Time	Date of the first data value in the series. To be programmatically updated if new records are added.	9/4/2003 7:00:00 AM	P
EndDateTime	Date/Time	Date of the last data value in the series. To be programmatically updated if new records are added.	9/4/2005 7:00:00 AM	P
BeginDateTimeUTC	Date/Time	Date of the first data value in the series in UTC. To be programmatically updated if new records are added.	9/4/2003 2:00 PM	P
EndDateTimeUTC	Date/Time	Date of the last data value in the series in UTC. To be programmatically updated if new records are added.	9/4/2003 2:00 PM	P
ValueCount	Integer	The number of data values in the series identified by the combination of the SiteID, VariableID, MethodID, SourceID and QualityControlLevelID fields. To be programmatically updated if new records are added.	50	P

Table: Sites

The Sites table provides information giving the spatial location at which data values have been collected.

Field Name	Data Type	Description	Example	Constraint	Default Value
SiteID	Integer Identity	Unique identifier for each sampling location.	37	M Unique Primary key	
SiteCode	Text (50)	Code used by organization that collects the data to identify the site	"10109000" (USGS Gage number)	M Unique Allows only characters in the range of A-Z (case insensitive), 0-9, and ".", "-", and "_".	
SiteName	Text (255)	Full name of the sampling site.	"LOGAN RIVER ABOVE STATE DAM, NEAR LOGAN,UT"	M Cannot contain tab, line feed, or carriage return characters	
Latitude	Real	Latitude in decimal degrees.	45.32	M ($>=$ -90 AND $<=$ 90)	
Longitude	Real	Longitude in decimal degrees. East positive, West negative.	-100.47	M ($>=$ -180 AND $<=$ 360)	
LatLongDatumID	Integer	Identifier that references the Spatial Reference System of the latitude and longitude coordinates in the SpatialReferences table.	1	M Foreign key	0 = Unknown
Elevation_m	Real	Elevation of sampling location (in m). If this is not provided it needs to be obtained programmatically from a DEM based on location information.	1432	O	NULL
VerticalDatum	Text (255)	Vertical datum of the elevation. Controlled Vocabulary from VerticalDatumCV.	"NAVD88"	O Foreign key	NULL
LocalX	Real	Local Projection X coordinate.	456700	O	NULL
LocalY	Real	Local Projection Y Coordinate.	232000	O	NULL

Field Name	Data Type	Description	Example	Constraint	Default Value
LocalProjectionID	Integer	Identifier that references the Spatial Reference System of the local coordinates in the SpatialReferences table. This field is required if local coordinates are given.	7	O Foreign key	NULL
PosAccuracy_m	Real	Value giving the accuracy with which the positional information is specified in meters.	100	O	NULL
State	Text (255)	Name of state in which the monitoring site is located.	"Utah"	O Cannot contain tab, line feed, or carriage return characters	NULL
County	Text (255)	Name of county in which the monitoring site is located.	"Cache"	O Cannot contain tab, line feed, or carriage return characters	NULL
Comments	Text (unlimited)	Comments related to the site.		O	NULL

The following rules and best practices should be followed when populating this table:

1. The SiteID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
2. The SiteCode field must contain a text code that uniquely identifies each site. The values in this field should be unique and can be an alternate key for the table. SiteCodes cannot contain any characters other than A-Z (case insensitive), 0-9, period ".", dash "-", and underscore "_".
3. The LatLongDatumID must reference a valid SpatialReferenceID from the SpatialReferences controlled vocabulary table. If the datum is unknown, a default value of 0 is used.
4. If the Elevation_m field is populated with a numeric value, a value must be specified in the VerticalDatum field. The VerticalDatum field can only be populated using terms from the VerticalDatumCV table. If the vertical datum is unknown, a value of "Unknown" is used.
5. If the LocalX and LocalY fields are populated with numeric values, a value must be specified in the LocalProjectionID field. The LocalProjectionID must reference a valid SpatialReferenceID from the SpatialReferences controlled vocabulary table. If the spatial reference system of the local coordinates is unknown, a default value of 0 is used.

Table: Sources

The Sources table lists the original sources of the data, providing information sufficient to retrieve and reconstruct the data value from the original data files if necessary.

Field Name	Data Type	Description	Example	Constraint	Default Value
SourceID	Integer Identity	Unique integer identifier that identifies each data source.	5	M Unique Primary key	
Organization	Text (255)	Name of the organization that collected the data. This should be the agency or organization that collected the data, even if it came out of a database consolidated from many sources such as STORET.	"Utah Division of Water Quality"	M Cannot contain tab, line feed, or carriage return characters	
SourceDescription	Text (unlimited)	Full text description of the source of the data.	"Text file retrieved from the EPA STORET system indicating data originally from Utah Division of Water Quality"	M	
SourceLink	Text (500)	Link that can be pointed at the original data file and/or associated metadata stored in the digital library or URL of data source.		O	NULL
ContactName	Text (255)	Name of the contact person for the data source.	"Jane Adams"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
Phone	Text (255)	Phone number for the contact person.	"435-797-0000"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
Email	Text (255)	Email address for the contact person.	"Jane.Adams@ dwq.ut"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
Address	Text (255)	Street address for the contact person.	"45 Main Street"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"

Field Name	Data Type	Description	Example	Constraint	Default Value
City	Text (255)	City in which the contact person is located.	"Salt Lake City"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
State	Text (255)	State in which the contact person is located. Use two letter abbreviations for US. For other countries give the full country name.	"UT"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
ZipCode	Text (255)	US Zip Code or country postal code.	"82323"	M Cannot contain tab, line feed, or carriage return characters	"Unknown"
Citation	Text (unlimited)	Text string that give the citation to be used when the data from each source are referenced.	"Data collected by USU as part of the Little Bear River Test Bed Project"	M	"Unknown"
MetadataID	Integer	Integer identifier referencing the record in the ISOMetadata table for this source.	5	M Foreign key	0 = Unknown or uninitialized metadata

The following rules and best practices should be followed when populating this table:

1. The SourceID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
2. The Organization field should contain a text description of the agency or organization that created the data.
3. The SourceDescription field should contain a more detailed description of where the data was actually obtained.
4. A default value of "Unknown" may be used for the source contact information fields in the event that this information is not known.
5. Each source must be associated with a metadata record in the ISOMetadata table. As such, the MetadataID must reference a valid MetadataID from the ISOMetadata table. The ISOMetatadata table should be populated with an appropriate record prior to adding a source to the Sources table. A default MetadataID of 0 can be used for a source with unknown or uninitialized metadata.
6. Use the Citation field to record the text that you would like others to use when they are referencing your data. Where available, journal citations are encouraged to promote the correct crediting for use of data.

Table: SpatialReferences

The SpatialReferences table provides information about the Spatial Reference Systems used for latitude and longitude as well as local coordinate systems in the Sites table. This table is a controlled vocabulary.

Field Name	Data Type	Description	Example	Constraint
SpatialReferenceID	Integer Identity	Unique integer identifier for each Spatial Reference System.	37	M Unique Primary key
SRSID	Integer	Integer identifier for the Spatial Reference System from http://www.epsg.org/.	4269	O
SRSName	Text (255)	Name of the Spatial Reference System.	"NAD83"	M Cannot contain tab, line feed, or carriage return characters
IsGeographic	Boolean	Value that indicates whether the spatial reference system uses geographic coordinates (i.e. latitude and longitude) or not.	"True", "False"	O
Notes	Text (unlimited)	Descriptive information about the Spatial Reference System. This field would be used to define a non-standard study area specific system if necessary and would contain a description of the local projection information. Where possible, this should refer to a standard projection, in which case latitude and longitude can be determined from local projection information. If the local grid system is non-standard then latitude and longitude need to be included too.		O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: SpeciationCV

The SpeciationCV table contains the controlled vocabulary for the Speciation field in the Variables table.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for Speciation.	"P"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of Speciation controlled vocabulary term. The definition is optional if the term is self explanatory.	"Expressed as phosphorus"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: TopicCategoryCV

The TopicCategoryCV table contains the controlled vocabulary for the ISOMetaData topic categories.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for TopicCategory.	"InlandWaters"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of TopicCategory controlled vocabulary term. The definition is optional if the term is self explanatory.	"Data associated with inland waters"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: Units

The Units table gives the Units and UnitsType associated with variables, time support, and offsets. This is a controlled vocabulary table.

Field Name	Data Type	Description	Example	Constraint
UnitsID	Integer Identity	Unique integer identifier that identifies each unit.	6	M Unique Primary key
UnitsName	Text (255)	Full text name of the units.	"Milligrams Per Liter"	M Cannot contain tab, line feed, or carriage return characters
UnitsType	Text (255)	Text value that specifies the dimensions of the units.	"Length" "Time" "Mass"	M Cannot contain tab, line feed, or carriage return characters
UnitsAbbreviation	Text (255)	Text abbreviation for the units.	"mg/L"	M Cannot contain tab, line feed, or carriage return characters

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: ValueTypeCV

The ValueTypeCV table contains the controlled vocabulary for the ValueType field in the Variables and SeriesCatalog tables.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for ValueType.	"Field Observation"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of the ValueType controlled vocabulary term. The definition is optional if the term is self explanatory.	"Observation of a variable using a field instrument"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: VariableNameCV

The VariableName CV table contains the controlled vocabulary for the VariableName field in the Variables and SeriesCatalog tables.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for Variable names.	"Temperature", "Discharge", "Precipitation"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of the VariableName controlled vocabulary term. The definition is optional if the term is self explanatory.		O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Table: Variables

The Variables table lists the full descriptive information about what variables have been measured.

Field Name	Data Type	Description	Example	Constraint	Default Value
VariableID	Integer Identity	Unique integer identifier for each variable.	6	M Unique Primary key	
VariableCode	Text (50)	Text code used by the organization that collects the data to identify the variable.	"00060" used by USGS for discharge	M Unique Allows only characters in the range of A-Z (case insensitive), 0-9, and ".", "-", and "_".	
VariableName	Text (255)	Full text name of the variable that was measured, observed, modeled, etc. This should be from the VariableNameCV controlled vocabulary table.	"Discharge"	M Foreign key	
Speciation	Text (255)	Text code used to identify how the data value is expressed (i.e., total phosphorus concentration expressed *as P*). This should be from the SpeciationCV controlled vocabulary table.	"P", "N", "NO3"	M Foreign key	"Not Applicable"
VariableUnitsID	Integer	Integer identifier that references the record in the Units table giving the units of the data values associated with the variable.	4	M Foreign key	
SampleMedium	Text (255)	The medium in which the sample or observation was taken or made. This should be from the SampleMediumCV controlled vocabulary table.	"Surface Water" "Sediment" "Fish Tissue"	M Foreign key	"Unknown"
ValueType	Text (255)	Text value indicating what type of data value is being recorded. This should be from the ValueTypeCV controlled vocabulary table.	"Field Observation" "Laboratory Observation" "Model Simulation Results"	M Foreign key	"Unknown"
IsRegular	Boolean	Value that indicates whether the data values are from a regularly sampled time series.	"True" "False"	M	"False"
TimeSupport	Real	Numerical value that indicates the time support (or temporal footprint) of the data values. 0 is used to indicate data values that are instantaneous. Other values indicate the time over which the data values are implicitly or explicitly averaged or aggregated.	0, 24	M	0 = Assumes instantaneous samples where no other information is available

Field Name	Data Type	Description	Example	Constraint	Default Value
TimeUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the time support. If TimeSupport is 0, indicating an instantaneous observation, a unit needs to still be given for completeness, although it is somewhat arbitrary.	4	M Foreign key	103 = hours
DataType	Text (255)	Text value that identifies the data values as one of several types from the DataTypeCV controlled vocabulary table.	"Continuous" "Sporadic" "Cumulative" "Incremental" "Average" "Minimum" "Maximum" "Constant Over Interval" "Categorical"	M Foreign key	"Unknown"
GeneralCategory	Text (255)	General category of the data values from the GeneralCategoryCV controlled vocabulary table.	"Climate" "Water Quality" "Groundwater Quality"	M Foreign key	"Unknown"
NoDataValue	Real	Numeric value used to encode no data values for this variable.	-9999	M	-9999

The following rules and best practices should be followed when populating this table:

1. The VariableID field is the primary key, must be a unique integer, and cannot be NULL. This field should be implemented as an auto number/identity field.
2. The VariableCode field must be unique and serves as an alternate key for this table. Variable codes can be arbitrary, or they can use an organized system. VaraibleCodes cannot contain any characters other than A-Z (case insensitive), 0-9, period ".", dash "-", and underscore "_".
3. The VariableName field must reference a valid Term from the VariableNameCV controlled vocabulary table.
4. The Speciation field must reference a valid Term from the SpeciationCV controlled vocabulary table. A default value of "Not Applicable" is used where speciation does not apply. If the speciation is unknown, a value of "Unknown" can be used.
5. The VariableUnitsID field must reference a valid UnitsID from the UnitsTable controlled vocabulary table.
6. Only terms from the SampleMediumCV table can be used to populate the SampleMedium field. A default value of "Unknown" is used where the sample medium is unknown.
7. Only terms from the ValueTypeCV table can be used to populate the ValueType field. A default value of "Unknown" is used where the value type is unknown.
8. The default for the TimeSupport field is 0. This corresponds to instantaneous values. If the TimeSupport field is set to a value other than 0, an appropriate TimeUnitsID must be specified. The TimeUnitsID field can only reference valid UnitsID values from the Units controlled vocabulary table. If the TimeSupport field is set to 0, any time units can be used (i.e., seconds, minutes, hours, etc.), however a default value of 103 has been used, which corresponds with hours.

9. Only terms from the DataTypeCV table can be used to populated the DataType field. A default value of "Unknown" can be used where the data type is unknown.
10. Only terms from the GeneralCategoryCV table can be used to populate the GeneralCategory field. A default value of "Unknown" can be used where the general category is unknown.
11. The NoDataValue should be set such that it will never conflict with a real observation value. For example a NoDataValue of -9999 is valid for water temperature because we would never expect to measure a water temperature of -9999. The default value for this field is -9999.

Table: VerticalDatumCV

The VerticalDatumCV table contains the controlled vocabulary for the VerticalDatum field in the Sites table.

Field Name	Data Type	Description	Examples	Constraint
Term	Text (255)	Controlled vocabulary for VerticalDatum.	"NAVD88"	M Unique Primary key Cannot contain tab, line feed, or carriage return characters
Definition	Text (unlimited)	Definition of the VerticalDatum controlled vocabulary. The definition is optional if the term is self explanatory.	"North American Vertical Datum of 1988"	O

This table is pre-populated within the ODM. Changes to this controlled vocabulary can be requested at http://water.usu.edu/cuahsi/odm/.

Appendix B. Data Versioning Within ODM

The main text of this document focuses on how ODM is structured to store observations data. It does not address how to manage editing data stored within ODM. Software applications based on ODM will have functionality that will allow data managers and database administrators to modify, delete, change, or otherwise make edits to data stored within ODM. In addition, these software tools will provide functionality to create derived datasets, or datasets that are calculated or derived from data already stored in ODM (i.e., calculate a time series of discharge from a time series of stage, or calculate a time series of daily average temperature from a time series of hourly observations). The purpose of this appendix is to clarify how data editing and versioning can be managed within the ODM schema.

Data Series Defined

In order to fully grasp the concepts that follow, the idea of a "data series" in the context of ODM must be clarified. A "data series" is an organizing principle that is present in the ODM. A data series consists of all of the data values associated with a unique site, variable, method, source, and quality control level combination. An example of the full specification for a data series is: "all of the raw unchecked (QualityControlLevel) water temperature (Variable) values measured in the Logan River near Logan, UT (Site) using a field temperature sensor (method) by Utah State University (Source)." Each record in the SeriesCatalog table of ODM represents a unique data series.

Rules for Editing and Deriving Data Series in ODM

The following rules are suggested so that versioning of and edits to data series can be managed within the ODM schema. Software applications that work with ODM should follow these rules. These rules are based on the default set of Quality Control Levels that are distributed with the ODM blank schema.

1. *Data versioning should be done at the data series level* – Within ODM, the concept of data versioning is related to the quality control level. Quality control level is a data series level attribute, and as such, changes to the quality control level should occur at the data series level rather than at the individual value level. For example, if an investigator wished to create a quality controlled Level 1 data series from a raw Level 0 data series, he/she should first make a copy of the raw Level 0 data series and then perform any edits and adjustments required in the quality control process to the copy. The edited copy then becomes the Level 1 data series, and the Level 0 data series is preserved intact.
2. *Data series with a QualityControlLevelCode of 0 cannot be edited* – Level 0 data series represent raw data from sensors (i.e., stage measured by a water level recorder) or other products derived from raw data (i.e., discharge that is programmatically derived from stage before the stage values have been quality controlled). By definition, Level 0 data have not been quality controlled and may contain significant errors and bad values. However, Level 0 data series represent the source from which all other derived data series are based, and as such should remain intact for archive purposes. Level 0 data series should not be used for analysis unless no other adequate options are available, and only if the user is aware that the data are raw. Level 0 data series can be removed entirely from the database, but only by removing the entire data series.
3. *Only one QualityControlLevel 0 data series can exist for a Site, Variable, and Method combination* – Only one raw data series for a Site, Variable, and Method combination can exist within an ODM database. If multiple sensors are measuring the same variable at the same site, the method description would have to distinguish between the two.
4. *Only one QualityControlLevel 1 data series can exist for each Site, Variable, and Method combination* – Once a Level 0 data series has been loaded to the database, a Level 1 data series

can be "derived" from that Level 0 data series. This is done by first making a copy of the Level 0 data series, second changing the QualityControlLevel of the copy to 1, and last doing any necessary filtering or editing required so that the Level 1 data series is acceptable as quality controlled. In most cases, the majority of the values within a Level 0 data series and its corresponding Level 1 data series will remain the same. However, where instruments malfunction or other conditions are present that affect the raw data values, Level 0 values may be deleted, adjusted, or otherwise edited in creating the Level 1 data series.

5. *Any edits to a data series are saved to that data series* – Level 0 data cannot be edited. With Levels 1 or higher, however, software applications should be allowed to edit and delete values. Each time an edit is made, the result should overwrite the previous value within a data series. In other words, edits should not create new data series, they should modify an existing one. This will be true even where edits are done within multiple editing sessions. The editing software should record the method or basis for any data edits in appropriate method records.

6. *Data series of Level 2 or higher can only be created from data series of Level 1 or higher* – Derived data series of Level 2 or higher can only be created from data series of Level 1 or higher. If a user wishes to create a derived data series from a Level 0 data series (such as discharge from raw, unchecked stage values) that derived data series would also be Level 0.

This page left intentionally blank.

APPENDIX C: DESCRIPTION OF ESF DATA LOADED TO THE ODM

Sources

<u>Loaded:</u>   Sources have been successfully loaded to the ESF DB.
<u>Comments:</u>   Contacts had to be updated.
<u>Total:</u> Four sources

Sites

<u>Loaded:</u>   Sites have been successfully loaded to the ESF DB.
<u>Comments:</u>   Sites had to be added. Elevations have been updated (Google Earth).
<u>Total:</u> 128 Sites.

SiteCode	Sitename	Latitude	Longitude	Elevation_m
UHL	Upper Hall Run at Clough Pike	39.08526	-84.295	270.34
SLT	South Lucy Tributary at Apple Rd	39.04119	-84.2129	254.8
SHC	Shaylor Crossing at County Lift Station in Shaylor Crossing Subdivision	39.08114	-84.2319	249.92
HST	Heiserman Stream at Rt 50 crossing and Experimental Stream Facility Intake	39.14839	-84.2565	159.4
KRT	Tributary to Kain Run at rt 276 crossing	39.06425	-84.0719	262.72
SHA	South Harsha Tributary at Bantam Rd crossing State Park	39.0072	-84.1368	248.4
AVR	Aveys Run at Cincinnati Nature Center's "Stream Site B"	39.12321	-84.2463	185
OWT	Owensville Tributary at County lift station off of St. Louis Rd	39.12775	-84.1364	243.52
USR	Upper Salt Run at County lift station off of Shepard Rd.	39.11801	-84.2596	191.1
HWR	Howard Run at Biddings Rd Crossing	39.12367	-84.0077	263.64
FMR	Fourmile Run at Elick Rd crossing	39.05586	-84.1681	182.26
HLR	Hall Run at Roundbottom Rd crossing	39.14039	-84.259	158.49
LRC	Lucy Run at Steve Wilson's property off of SR 222	39.05701	-84.1795	182.57
GRR	Grassy Fork at Glancey-Marathon Rd crossing	39.13286	-84.0152	262.12
SHR	Shaylor Run at Bethel Rd Crossing	39.11834	-84.2163	169.16
CLC	Cloverlick Creek in State Park at Austin Rd off of 133	38.98984	-84.0594	244.54
STC	Stonelick Creek at RT 50 crossing	39.12203	-84.1986	163.97
EFG	East Fork River @ Morgan Rd	39.2191	-83.9156	279.18
EFY	East Fork River @ Fayetteville	39.18764	-83.9381	27.95
EFM	East Fork River @ Marathon	39.13799	-84.0029	261.2
EFB	East Fork River @ Blue Sky Parkway	39.11526	-84.0246	258.46
ELI	East Fork River @ Williamsburg	39.05272	-84.0515	240.78
DWT	Drinking Water Treatment Plant	39.05131	-84.1361	212.74
DAM	East Fork River Below East Fork Lake Dam	39.02614	-84.1482	186.53

EFK	East Fork River @ Experimental Stream Facility	39.14608	-84.2519	154.53
EFC	East Fork Lake @ Confluence w/ Little Miami River	39.15574	-84.2884	150.56
ESF	Experimental Stream Facility Weather Station	39.14723	-84.2547	159.71
LEF	Lower East Fork Waste Water Treatment Plant Effluent	39.14676	-84.2559	158.18
2EFR11205	Poplar Creek and Ohio 232, Bethel O. Quad.	38.96139	-84.1067	253.58
2EFR21001	East Fork Lake, Cloverlick Inflow Tributary above Poplar Creek	38.99917	-84.0858	219.14
2EFR10002	EFLMR at Ohio 32 Bridge at Williamsburg, USGS Gage, Williamsburg, O. Quad (Equivalent to EPA "ELI" and CCOEQ "EFRM34.8")	39.0525	-84.0506	240.78
2EFR20004	East Fork Lake, EFLMR Main Inflow Tributary, Williamsburg, O. Quad	39.01999	-84.1311	216.09
2EFR20022	East Fork Lake, New Site 2010 to account for mixing of major inflows	39.01722	-84.1017	207.56
2EFR20023	East Fork Lake, New Site 2010 upstream point of 'narrows'	39.01611	-84.1114	215.18
2EFR20024	East Fork Lake, at mile 24. 0.5mi S of Elk Lick, Ohio, ON Batavia, OH Quad. At cross section from STA 20024 to mouth of Slabcamp Run, 2/10 of length from left bank	39.02	-84.1311	199.94
2EFR20025	East Fork Lake, New Site 2010, Cove section North of STA 2EFR20024	39.02833	-84.1319	195.06
2EFR20001	Located on East Fork Lake at the Log Boom. (Lake Side of "DAM" near USACE Flow Control Structure)	39.02917	-84.0867	191.41
2EFR10000	EFLMR at CO. RD. Bridge 0.1 Miles off Elk Lick Road, Batavia, O. Quad. (Near EPA "DAM" Site)	39.02611	-84.1478	186.22
EFL	East Fork Lake near Bob McEwen Drinking Water Treatment Plant Intake (Sampled from bridge to intake structure)	39.03695	-84.1387	212.74
CEC	Cemetery Creek	39.08222	-84.1767	193.54
1IJ00060001	Ohio Asphaltic Limestone Corp; Final effluent discharge from sedimentation basin to an unnamed tributary of Turtle Creek	39.24972	-83.6814	354.77
1IJ00060002	Ohio Asphaltic Limestone Corp; Final effluent discharge from sedimentation basin to an unnamed tributary of Dodson Creek	39.24306	-83.6825	338.92
1IN00287001	Hanson Aggregates Midwest - Highland Quarry Industrial Discharge Facility	39.19417	-83.7222	315.45
1PB00105001	Village of Lynchburg Municipal WWTP, Final Effluent	39.23806	-83.8003	297.47
1PB00105801	Village of Lynchburg Municipal WWTP, Upstream Monitoring (Pearl St. Covered Bridge)	39.24306	-83.7953	299.3

1PB00105 901	Village of Lynchburg Municipal WWTP, Downstream Monitoring	39.2375	-83.8006	297.17
1PG00100 001	Rolling Acres Municipal WWTP Final effluent discharge to an unnamed tributary of Dodson Creek. Monitoring just after dechlorination.	39.22222	-83.6856	344.41
2EFR1120 4	Poplar Creek and St Route125	38.97379	-84.1058	247.5
2EFR2000 5	East Fork Lake near combined confluence of Cloverlick and Poplar Creek's inflows	39.02336	-84.1013	218.8
2EFR2100 0	East Fork Lake near Cabin Run Inflow	39.01527	-84.0984	208.2
2EFR2300 1	East Fork Lake near Slabcamp Run Inflow	39.03919	-84.1364	212.4
5MILECR 0.5	Five Mile Creek @ Bluesky; upstream from Bluesky Pkwy	39.1136	-84.0203	264.9
BARNS1. 9	Barnes Run @ Bethel-Concord; Bethel Concord Rd	39.0082	-84.0733	249.92
BRUSH0. 3	Brushy Fork RM 0.3 at Titus Road	39.143	-84.1447	199.02
CABIN1.5	Cabin Run; At Campground Loop G	39.0327	-84.1019	237.7
CLOV5.1	Clover Creek RM 1.9 at St. Rt. 133	38.9856	-84.0528	250.84
CWL	Cornwell Farm, stream crossing at driveway to Stahl's Property	39.18389	-84.0133	286.19
DODSN0. 1	Dodson Creek RM 0.1 at mouth; Dodson Creek @ Crampton Rd.	39.22285	-83.8127	294.12
DWT-0	Inside Bob McEwen Drinking Water Treatment Plant, Pumped from inside intake structure on lake	39.05169	-84.1356	261.2
DWT-PCD	Drinking Water Post Chlorine Disinfection, before discharge to distribution system	39.05169	-84.1356	262.4
DWT-PFL	Effluent of all filters currently operating	39.05169	-84.1356	262.1
DWT-PMN	Drinking Water Treatment Plant - Post Manganese addition before coagulant is added	39.05169	-84.1356	261.5
DWT-SET	DWT- effluent from settling basin	39.05169	-84.1356	261.8
E01	Mesocosm 1 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E02	Mesocosm 2 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E03	Mesocosm 3 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E04	Mesocosm 4 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E05	Mesocosm 5 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7

E06	Mesocosm 6 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E07	Mesocosm 7 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
E08	Mesocosm 8 in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
EFRM0.7	EFLMR RM 0.7 at S. Milford Rd. (Same as EPA EFC site 0)	39.1533	-84.2975	151.48
EFRM15.6	East Fork @ State Route 222	39.0624	-84.1789	176.47
EFRM34.8	EFLMR @ Main st. autosampler; Williamsburg Main Street Bridge	39.0525	-84.0504	240.5
EFRM44.1	EFLMR RM 44.1 at Blue Sky Park Rd.	39.1158	-84.025	259.98
EFRM60.6	EFLMR RM 60.6 at US 50 Fayetteville Lift Station	39.1865	-83.9372	271.56
EFRM70.1	East Fork @ Wise Road; Clinton Co Wise Road Bridge	39.22513	-83.8255	298.38
EFRM75.3	East Fork @ Canada Rd; Canada Road Bridge	39.2731	-83.7812	303.87
EFRM9.1	East Fork @Stonelick-Olive Branch; upstream Stonelick-Olive branch bridge	39.11869	-84.2086	163.06
ERS	East Fork River Supply Monitoring Station in the Experimental Stream Facility, 1003 US Route 50, Milford, OH 45150	39.14718	-84.2548	159.7
GRSSY0.2	Grassy Fork @ GC-Marathon Rd autosampler; Glancy Corner-Marathon Road	39.1329	-84.0153	262.1
GRSSY3.0	Grassy Fork Marathon Edenton; 5764 Marathon-Edenton Road	39.1664	-84.0092	274.31
GRSSY3.2	Grassy Fork @ 131 Brown County; river left tributary, U.S. 131, Brown County	39.1741	-83.9982	278.88
HALL0.2	Hall Run @ Roundbottom Rd	39.1401	-84.2593	158.8
HWRD0.4	Howard Run @ Burdsall; Burdsall Road	39.1239	-84.0073	264.6
NLT	North Lucy Tributary at confluence with South Lucy Tributary	39.04121	-84.212	250.5
NWT	Cedarville Rd stream crossing north of Main St., Newtonsville OH	39.185	-84.095	265.47
PLEAS0.2	Pleasant Run Hutchinson Prop; Hutchinson Road off SR 133	39.1084	-84.0374	258.8
POPLR2.1	Poplar Creek @ Macedonia; Macedonia Road Bridge	38.982	-84.101	237.4
SHYLR1.7	Shayler Run @ Baldwin Rd	39.1183	-84.2163	169.2
SLAB0.5	Slabcamp Run accessed via Greenbriar Rd; trailhead across from NE WTP pond	39.0511	-84.129	245.7
ST13.4	Stonelick Creek RM 13.4 at 727	39.2097	-84.0889	262.42

ST17.3	Stonelick Creek @ Stonelick Tr; u/s from Stonelick Tr Private Drive	39.232	-84.044	268.2
ST5.7	Stonelick Creek RM 0.9 at Anstaett Rd.	39.1429	-84.1492	195.67
STEFLMR	Stonelick Creek at US 50	39.1222	-84.1991	164
TBS	Twin Bridges Road, stream crossing just east of horse camp	39.03806	-84.1203	252.97
ULREY1.3	Ulrey Run@ St. Rt. 125; u/s from St. Rt. 125	39.0017	-84.1515	239.3
1IN00116601	USEPA Experimental Stream Fac	39.1475	-84.2553	160
1IN00123902	CECOS International Inc	39.12694	-84.0519	271
1PA00005001	New Vienna WWTP	39.32889	-83.7033	336
1PB00001001	Batavia WWTP	39.08222	-84.1767	170
1PB00034001	Williamsburg WWTP	39.05667	-84.0464	241
1PC00005001	Milford STP	39.16444	-84.2836	153
1PD00024001	Fayetteville Perry Twp WWTP	39.18361	-83.9392	272
1PH00031001	Martinsville-Midland WWTP	39.32333	-83.8694	297
1PK00010001	Middle East Fork Regional WWTP	39.08861	-84.1872	170
1PN00000002	US DOA William H Harsha Lake	39.02639	-84.145	249
1PP00020001	Stonelick State Park Campgrounds	39.22444	-84.0578	269
1PR00116001	Cincinnati Nature Center - Rowe Woods Recreation	39.12778	-84.2472	233
1PT00077001	Clermont NE Local Schools WWTP	39.12806	-84.1072	256
1PV00002001	Holly Towne MHP	39.00722	-84.1783	262
1PV00009001	Orchard Lake MHP	39.19111	-84.2464	246
1PV00034001	Forest Creek MHP	38.99083	-84.1589	255
1PV00074001	Royal Hills MHP	39.18	-84.2511	232
1PX00059001	Locust Ridge Nursing Home Inc	38.9875	-84.0225	274
1PZ00029001	Snow Hill Country Club	39.34917	-83.7158	335
1PN00000001	US DOA William H Harsha Lake	39.22	-84.06	272
2PD00024002	South Pleasant Street Lift Station Combined Sewer Overflow	40	-85	-9999

DAMCS	Located on East Fork Lake at the Log Boom. (Lake Side of ""DAM"" near USACE Flow Control Structure)	39.02	-84.1536	NULL
DRI	Site managed by Jake Beaulieu	40	-85	NULL
DWF	Drinking Water Treatment Plant, Post Filtration, Sampled by Macke for Elovitz	39.0369	-84.1387	NULL
GWS	ESF Ground Water Source	39.14718	-84.2548	NULL
REF	Site managed by Jake Beaulieu	40	-85	NULL
ROA	Site managed by Jake Beaulieu	40	-85	NULL
TMC	Site managed by Jake Beaulieu	40	-85	NULL
QAC	Quality Assurance Quality Control Sample	39.14718	-81.2548	NULL
GWT	Grailville Treatment Wetland	40	-85	NULL
FF	??	40	-85	NULL

Physico-chemical Data Water and Sediment - EXCEL Sheets

Variables

Loaded:

The variables listed below have been loaded and mapped to the existing ODM vocabulary. VariableCode is a unique identifier and consists of the AnalyteName-Matrix.

DRP-SW	Phosphorus, orthophosphate	P
NH4-SW	Nitrogen, NH4	N
NH4-Urease-SW	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NO2-SW	Nitrogen, nitrite (NO2) nitrogen	N
NO2-3-SW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TDN-SW	Nitrogen, total dissolved	N
TDP-SW	Phosphorus, total dissolved	P
TN-SW	Nitrogen, total	N
TNH4-SW	Nitrogen, NH3 + NH4	N
TNO2-SW	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-3-SW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TOC-SW	Carbon, total organic	C
TP-SW	Phosphorus, total	P
TRP-SW	Phosphorus, orthophosphate dissolved	P
TUREA-SW	Urea, total	N
UREA-SW	Urea, dissolved (filtered)	N
DOC-SW	Carbon, dissolved organic	C
DOC-AD	Carbon, total organic	C
DOC-BP	Carbon, dissolved organic	C
DOC-DW	Carbon, dissolved organic	C
DOC-IG	Carbon, dissolved organic	C
DOC-WW	Carbon, dissolved organic	C
DRP-AD	Phosphorus, orthophosphate	P
DRP-BP	Phosphorus, orthophosphate	P
DRP-DW	Phosphorus, orthophosphate	P

DRP-IG	Phosphorus, orthophosphate	P
DRP-WW	Phosphorus, orthophosphate	P
NH4-AD	Nitrogen, NH4	N
NH4-BP	Nitrogen, NH4	N
NH4-DW	Nitrogen, NH4	N
NH4-IG	Nitrogen, NH4	N
NH4-Urease-AD	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NH4-Urease-BP	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NH4-Urease-DW	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NH4-Urease-IG	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NH4-Urease-WW	Ammonium analyzed as an endpoint to 24hr Urease Assay	N
NH4-WW	Nitrogen, NH4	N
NO2-3-AD	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
NO2-3-BP	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
NO2-3-DW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
NO2-3-IG	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
NO2-3-WW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
NO2-AD	Nitrogen, nitrite (NO2) nitrogen	N
NO2-BP	Nitrogen, nitrite (NO2) nitrogen	N
NO2-DW	Nitrogen, nitrite (NO2) nitrogen	N
NO2-IG	Nitrogen, nitrite (NO2) nitrogen	N
NO2-WW	Nitrogen, nitrite (NO2) nitrogen	N
TDN-AD	Nitrogen, total dissolved	N
TDN-BP	Nitrogen, total dissolved	N
TDN-DW	Nitrogen, total dissolved	N
TDN-IG	Nitrogen, total dissolved	N
TDN-WW	Nitrogen, total dissolved	N
TDP-AD	Phosphorus, total dissolved	P
TDP-BP	Phosphorus, total dissolved	P
TDP-DW	Phosphorus, total dissolved	P
TDP-IG	Phosphorus, total dissolved	P
TDP-WW	Phosphorus, total dissolved	P
TN-AD	Nitrogen, total	N
TN-BP	Nitrogen, total	N
TN-DW	Nitrogen, total	N
TNH4-AD	Nitrogen, NH3 + NH4	N
TNH4-BP	Nitrogen, NH3 + NH4	N
TNH4-DW	Nitrogen, NH3 + NH4	N
TNH4-IG	Nitrogen, NH3 + NH4	N
TNH4-WW	Nitrogen, NH3 + NH4	N
TN-IG	Nitrogen, total	N
TNO2-3-AD	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N

TNO2-3-BP	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TNO2-3-DW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TNO2-3-IG	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TNO2-3-WW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TNO2-AD	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-BP	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-DW	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-IG	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-WW	Nitrogen, nitrite (NO2) nitrogen	N
TN-WW	Nitrogen, total	N
TOC-AD	Carbon, total organic	C
TOC-BP	Carbon, total organic	C
TOC-DW	Carbon, total organic	C
TOC-IG	Carbon, total organic	C
TOC-WW	Carbon, total organic	C
TP-AD	Phosphorus, total	P
TP-BP	Phosphorus, total	P
TP-DW	Phosphorus, total	P
TP-IG	Phosphorus, total	P
TP-WW	Phosphorus, total	P
TRP-AD	Phosphorus, orthophosphate dissolved	P
TRP-BP	Phosphorus, orthophosphate dissolved	P
TRP-DW	Phosphorus, orthophosphate dissolved	P
TRP-IG	Phosphorus, orthophosphate dissolved	P
TRP-WW	Phosphorus, orthophosphate dissolved	P
TUREA-AD	Urea, total	N
TUREA-BP	Urea, total	N
TUREA-DW	Urea, total	N
TUREA-IG	Urea, total	N
TUREA-WW	Urea, total	N
UREA-AD	Urea, dissolved (filtered)	N
UREA-BP	Urea, dissolved (filtered)	N
UREA-DW	Urea, dissolved (filtered)	N
UREA-IG	Urea, dissolved (filtered)	N
UREA-WW	Urea, dissolved (filtered)	N
TNH4-SWNR	Nitrogen, NH3 + NH4	N
TUREA-SWNR	Urea, total	N
TN-SWNR	Nitrogen, total	N
TP-SWNR	Phosphorus, total	P
TUREA-GW	Urea, total	N
TRP-GW	Phosphorus, orthophosphate dissolved	P
TNH4-GW	Nitrogen, NH3 + NH4	N

TNO2-DI	Nitrogen, nitrite (NO2) nitrogen	N
TNO2-3-SWNR	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TNO2-SWNR	Nitrogen, nitrite (NO2) nitrogen	N
TRP-SWNR	Phosphorus, orthophosphate dissolved	P
TNO2-3-GW	Nitrogen, nitrite (NO2) + nitrate (NO3) nitrogen	N
TN-GW	Nitrogen, total	N
TP-GW	Phosphorus, total	P

Comments: 1. The variables in the Excel sheets had to be changed to enable them to be committed to the Database.
2. Specifications, VariableUnitsID and Value Type had to be checked.
3. VariableNameCV. The following new variable terms have been loaded to this table: urea, dissolved (filtered) and urea, total urea.
4. File: ESF-DataValues-final.csv

OffsetTypes
Loaded: The following new offset terms have been added

Tray 1.1A

Tray 1.2A

Tray 1.3A

Tray 1.1B

Tray 1.2B

Tray 1.3B

Tray 2.1A

Tray 2.2A

Tray 2.3A

Tray 2.1B

Tray 2.2B

Tray 2.3B

Tray 3.1A

Tray 3.2A

Tray 3.3A

Tray 3.1B

Tray 3.2B

Tray 3.3B

Tray 4.1A

Tray 4.2A

Tray 4.3A

Tray 4.1B

Tray 4.2B

Tray 4.3B

Tray 5.1A

Tray 5.2A

Tray 5.3A

Tray 5.1B

Tray 5.2B

Tray 5.3B

Center

Left

Right

Below Surface Water

Below Surface

Sensor Data Water-immersed Sensors for Physico-chemical Variables
http://66.161.146.122/ESF_Field_Data/

Friday,	July	08,	2011	1:34 PM	1549191	CEC.csv
Friday,	July	08,	2011	1:10 PM	880313	CLC.csv
Friday,	July	08,	2011	1:00 PM	74762	FMR.csv
Friday,	July	08,	2011	1:43 PM	14801139	HST.csv
Friday,	July	08,	2011	1:10 PM	476831	HWR.csv
Friday,	July	08,	2011	12:57 PM	4335011	KRT.csv
Friday,	July	08,	2011	1:00 PM	4382015	LRC.csv
Friday,	July	08,	2011	12:58 PM	3695128	OWT.csv
Friday,	July	08,	2011	12:58 PM	4439957	SHA.csv
Friday,	July	08,	2011	12:56 PM	4352695	SHC.csv
Friday,	July	08,	2011	1:02 PM	1482854	SHR.csv
Friday,	July	08,	2011	12:55 PM	3915199	SLT.csv
Friday,	July	08,	2011	12:55 PM	4415080	UHL.csv
Friday,	July	08,	2011	12:59 PM	3998020	USR.csv

Comments: For all the sensor data files, except for HST.csv, one new column has been added that combines date and time and a second column with OffsetTypeID (3 which describes Below Surface water in ft).

Variables
Loaded: The sensor data variables listed below have been loaded to the ODM.

VariableCode	VariableName	VariableUnitsID	ValueType	TimeSupport	TimeUnitsID	DataType	GeneralCategory
Temp	Temperature	96	Field Observation	10	104	Continuous	Water Quality
SPCOND	Specific conductance	192	Field Observation	10	104	Continuous	Water Quality
DoSat	Oxygen, dissolved percent of saturation	1	Field Observation	5	104	Average	Water Quality

C-10

DO	Oxygen, dissolved	199	Field Observation		5	104	Continuous	Water Quality
PH	pH	309	Field Observation		5	104	Continuous	Water Quality
ORP	Reduction potential	169	Field Observation		5	104	Continuous	Water Quality
TURB	Turbidity	221	Field Observation		5	104	Continuous	Water Quality

This page intentionally left blank

APPENDIX D: DESCRIPTION OF SHEPHERD's CREEK DATA LOADED TO THE ODM

Sources
Loaded: Sources have been successfully loaded to the Shepherd Creek DB.
Comments: Contacts had to be updated.
Total:

Sites
Loaded: Sites have been successfully loaded to the SC DB.
Comments: Sites had to be added. Elevations have been updated.
Total: 6 Sites.

SiteCode	Sitename	Latitude	Longitude	Elevation_m
CON	Con5(Catch)	39.17329	-84.5797	215
PWR	Pwr2(Sub1)	39.18222	-84.5814	242
DRI	Dri3(Sub2)	39.1773	-84.5785	225
ROA	Roa4(Sub3)	39.17879	-84.576	233
URB	Urb6(Sub4)	39.18235	-84.5746	242
REF7	Ref7(Sub5)	39.17649	-84.5781	223

Physico-chemical Data Water - Excel Sheet
Variables
Loaded:

The variables listed below have been loaded and mapped to the existing ODM vocabulary. VariableCode is a unique identifier and consists of the AnalyteName-Matrix.

TP	Phosphorus, total	P
Temp	Temperature	Not Applicable
TN	Nitrogen, total	N
Ecoli	E-coli	Not Applicable
Turb	Turbidity	Not Applicable
Alk	Alkalinity, carbonate	CaCO3
DOC	Carbon, dissolved organic	Unknown
TOC	Carbon, total organic	Unknown
Cl	Chloride	Cl
Br	Bromide	Br
NO3	Nitrogen, nitrate (NO3)	NO3
SO4	Sulfate	SO4
DIN	Nitrogen, dissolved inorganic	Unknown
TKN	Nitrogen, total kjeldahl	Unknown
NH3-N	Nitrogen, NH3	N
TDP	Phosphorus, total dissolved	Unknown
Al	Aluminum, dissolved	Al
Ca	Calcium	Ca
Cu	Copper, dissolved	Cu
Fe	Iron, dissolved	Fe
Mg	Magnesium	Mg

Mn	Manganese, dissolved	Mn
K	Potassium	K
Na	Sodium	Na
Zn	Zinc, dissolved	Zn
oPO4	Phosphorus, ortophosphate	PO4
SSC	Suspended Sediment Concentration	Not Applicable
ZnTR	Zinc, total reactive	Zn
MnTR	Manganese, total reactive	Mn
FeTR	Iron, total reactive	Fe
CuTR	Copper, total reactive	Cu
AlTR	Aluminum, total reactive	Al

Comments:
1. The variables in the Excel sheets had to be changed to enable them to be committed to the Database.
2. Specifications, VariableUnitsID and Value Type had to be checked.
3. VariableNameCV. The following new variable terms have been loaded to this table: Zinc, total reactive; Manganese, total reactive; Iron, total reactive; Copper, total reactive; Aluminum, total reactive.
4. File: ShepherdCreek-DataValues-final.csv

HYDROLOGY OF JACOB'S WELL SPRING

A tutorial for using HydroDesktop to discover and access water data

Presented at the University of Cincinnati

September 6, 2011

by:

Dr. Tim Whiteaker (twhit@mail.utexas.edu)
Center for Research in Water Resources
The University of Texas at Austin

And

Dr. David Tarboton
Utah State University

Distribution

Funding

Table of Contents

Introduction

CUAHSI-HIS enables sharing of water data

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information Systems project (CUAHSI-HIS) has devoted itself to improving access to time series of water data. Towards that end, CUAHSI-HIS has developed standards for sharing data that make it easier to ask for water data and interpret what comes back from a given data source. CUAHSI-HIS also maintains a catalog of data available from organizations that use CUAHSI-HIS standards, essentially serving as a search engine for water data. The result is a universal mechanism for accessing time series data, greatly simplifying the typically laborious task of getting the data you need to do your analysis. But how do people who are unfamiliar with CUAHSI-HIS standards use this system? That's where HydroDesktop comes in.

HydroDesktop uses CUAHSI-HIS to help you find water data

HydroDesktop is a free and open source Geographic Information Systems (GIS) application that helps you discover, use, and manage hydrologic time series data published with CUAHSI-HIS. It handles the details of how to work with CUAHSI-HIS so that you don't have to. HydroDesktop includes data query, download, visualization, graphing, analysis and modeling capabilities. The result is a spatially-enabled system that facilitates the aggregation of observational data describing the water environment.

Let's use HydroDesktop to learn about Jacob's Well Spring

This document presents an exercise that shows how to use HydroDesktop to find water data for Jacob's Well Spring in Texas. With some simple analysis, you will compare characteristics of this groundwater-dominated system with those of a nearby river. During the exercise, you will learn about some of the most commonly-used tools in HydroDesktop.

<u>**Related Links**</u>**:**
HydroDesktop - http://hydrodesktop.codeplex.com/
CUAHSI Hydrologic Information System - http://his.cuahsi.org/

About Jacob's Well Spring

The underwater cave known as Jacob's Well emerges in Hays County, Texas, at Jacob's Well Spring where it serves as one of the primary sources of water for Cypress Creek, which later flows into the Blanco River. The clear, crisp water cools down many Texans as it moves through the Blue Hole swimming area near Wimberley, Texas. This karst spring has been impacted in recent years by development in Hays County and increasing demands on the Middle Trinity Aquifer (Davidson, 2008).

Figure 0-1 Jacob's Well Spring (San Marcos Local News, 2009)

In 2005, a monitoring station was installed at Jacob's Well Spring 18 meters below the ground surface, reporting flow and temperature conditions at 15-minute intervals. The data for this station are accessible via the US Geological Survey's National Water Information System (USGS NWIS).
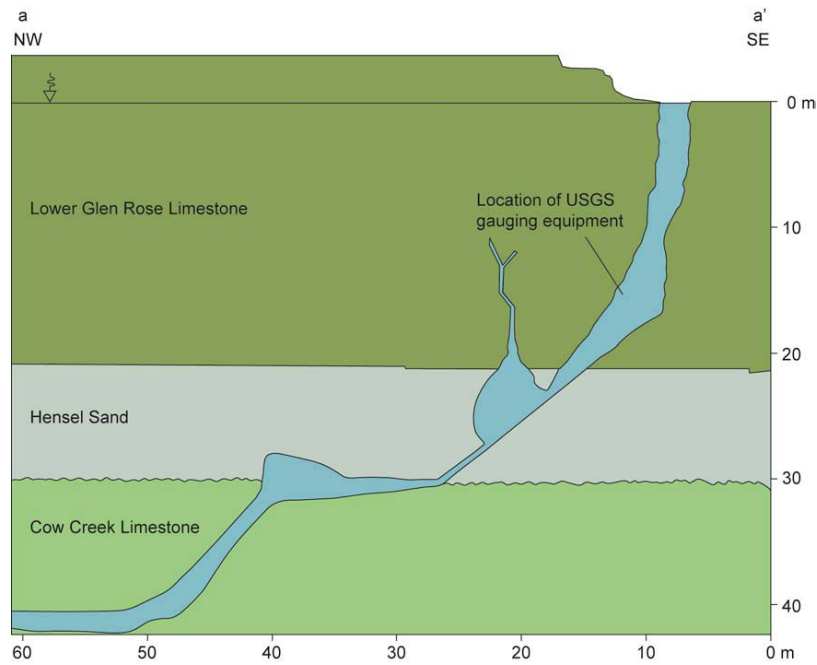
Figure 0-2 Cross-sectional diagram of Jacob's Well (Davidson, 2008)

Figure 0-3 Jacob's Well Spring Monitoring Station (United States Geological Survey, 2007)

For more information about the spring, please read the 2008 Masters thesis of Sarah Cain Davidson from The University of Texas at Austin.

Goals and Objectives

The goal of this exercise is to introduce you to the tools and functions available in HydroDesktop that allow you to search for and synthesize hydrologic time series data in an area of interest. This exercise will teach you how to find and obtain data for Jacob's Well Spring in Texas and compare data characteristics using the analysis capabilities of HydroDesktop.

Objectives for this exercise include:

- Find streamflow and temperature data for Jacob's Well Spring in Texas.
- Identify useful time series and download them.
- Visualize time series data in graphs.
- Export time series data for use in other programs.

Computer and Data Requirements

At the time of this writing, HydroDesktop is still in beta stages, and thus changes are made frequently to fix bugs and enhance the software. Therefore, it is recommended that you install the version of HydroDesktop that was used to prepare this exercise: **Version 1.2.591 Beta Release**. This version is only compatible with a Windows operating system such as **Windows XP or Windows 7**. You will also need an Internet connection since you will be accessing online resources to download time series data.

To install HydroDesktop:
1. In a Web browser, navigate to http://hydrodesktop.codeplex.com/.
2. Click the **Downloads** link near the top left of the page.
3. Find the link for the **1.2.601 Beta Release** installer and click it.
4. Read the license and agree to it.
5. Save and run the installer, accepting all defaults. The installer will guide you through the rest.

An advanced portion of the exercise involves using the R statistical package within HydroDesktop. R is a separate program from HydroDesktop, so you will need to install it if it is not already installed on your computer. It is not included with the HydroDesktop installation.

To install R:
1. In a Web browser, navigate to **http://www.r-project.org/**.
2. Click the **download R** link.
3. Click the link for a download site near your current location.
4. Click the link for your operating system (most likely Windows if you are using HydroDesktop).
5. Click the link for the **base** installation.
6. Click the link to **Download R**.
7. Run the setup file and follow the instructions to complete the installation.

Participating in the Open Source Community

HydroDesktop is an open source product, which means that anyone can see the source code used to create the program and contribute to its development. Even if you aren't a programmer, you can still participate in the discussion forums and post bugs or feature requests in the issue tracker.

The home for HydroDesktop is on CodePlex, a Web site for open source software. To add to the discussions or post a bug, you must first register for your free CodePlex account. Once you have a CodePlex account, you can log in at http://hydrodesktop.codeplex.com/ and start contributing. The community is really what drives open source software development, so this is an exciting opportunity to make your voice heard.

You are encouraged to provide feedback on any issue or problem you may encounter throughout this exercise. Feel free to utilize online resources such as the issue tracker on the HydroDesktop Web site when providing feedback. In this exercise you'll learn how to access these resources directly through HydroDesktop.

Exercise Procedure

Suppose you live in Hays County in Texas, and for years you have enjoyed taking a dip in the Blue Hole swimming area along Cypress Creek during hot Texas summers. As population growth and increased groundwater pumping threaten Jacob's Well Spring, the primary source of water for Cypress Creek, you decide to learn more about this valuable resource. In this exercise, you'll use HydroDesktop to find temperature data and see how it compares to a nearby river.

> **IMPORTANT**
>
> At the time of this writing, HydroDesktop is still in the beta stages of software development and thus still contains bugs. We are working hard to fix these bugs, but you may want to closely and carefully follow the exercise procedure in the meantime in order to minimize bugs encountered.

Getting To Know HydroDesktop

Let's open HydroDesktop and get to know its user interface.
1. Open HydroDesktop (Start | All Programs | CUAHSI HIS | HydroDesktop | HydroDesktop).
2. Choose to create a new **North America** project and click **OK**.

Take a moment to explore the user interface. As you can see, HydroDesktop looks much like a typical GIS interface. It supports complex layer symbologies, access to online map services, and custom programmed tools and plugins. It even comes with some basemap shapefile data which are already added to the map. What sets HydroDesktop apart from other GIS applications, is the ability to query for hydrologic time series data.

Notice that HydroDesktop presents many of its controls on a ribbon, much like modern versions of Microsoft Office. The ribbon is organized into tabs which contain groups of buttons and tools. There is also an orb for accessing basic functions like saving and printing.
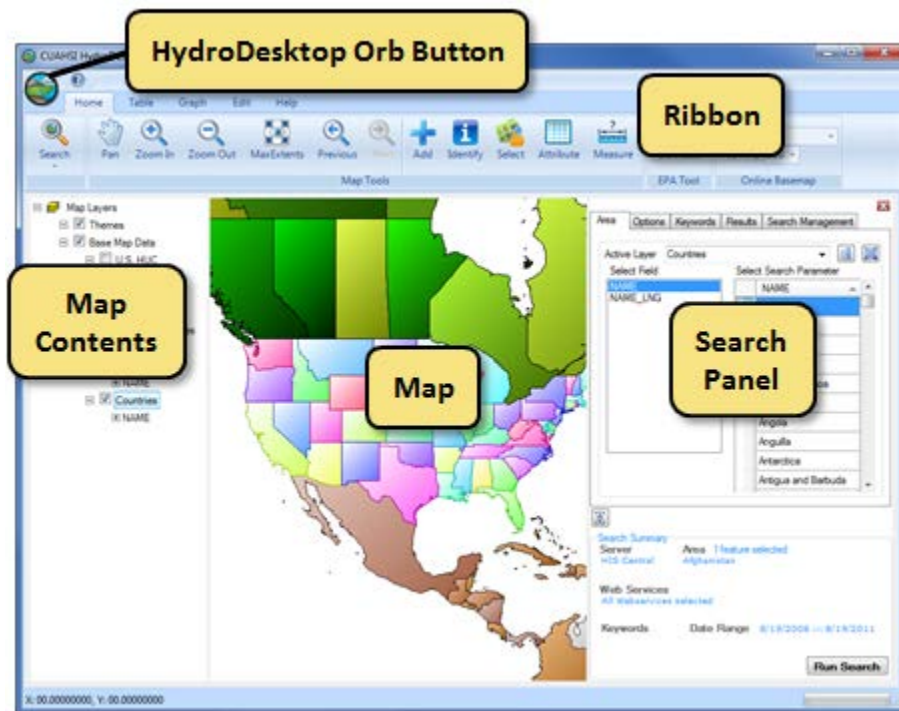
Figure 0-1 HydroDesktop User Interface

E-8

If you have comments or issues as you work through this exercise, you can find helpful resources on the Help tab. The buttons on this tab let you view documentation, jump to the discussion forums or issue tracker, email for help, or submit a comment.

3. Click the **Help** tab in the ribbon to view the buttons available on that tab.
4. Click the **Issues** button to open the issue tracker on the HydroDesktop Web site.

Figure 0-2 Using the Help Tab To Open the Issue Tracker

5. Close the Issue Tracker Web page.

Creating a Project

HydroDesktop manages your work within *projects*. A HydroDesktop project file (.hdprj) contains information about what geospatial layers you have in your map and how those layers are symbolized. These layers are stored in shapefiles, a widely available GIS data format. The shapefiles such as state boundaries that are included with HydroDesktop are located in its installation folder, e.g., C:\Program Files\CUAHSI HIS\HydroDesktop\maps\BaseData.

The HydroDesktop project file also connects your work to a database (.sqlite file) where temporal data are stored. This is where the time series data that you download through HydroDesktop are saved. A relational database is much more efficient at storing time series data than shapefiles, and HydroDesktop uses a free database called SQLite for this purpose.

You can create projects to organize your work, and you can save the project so that you can open it again later. When you first open HydroDesktop, it sets up a clean map and loads the default system database. In order to better manage the work in this exercise, you will give this project a name and save it somewhere meaningful to you.

To save the project and database:
1. Click the **HydroDesktop Orb** button.
2. In the **Orb** menu, click **Save Project**.
3. Choose a location to save your project such as your desktop.
4. Name the project **springs** and then click **Save**.

The text in the title bar of the HydroDesktop window should now include the name of your project. HydroDesktop has also created a database for your project named "springs.sqlilte." This database is saved in the same location as your project file. You are now ready to work within your newly saved project and database.

Searching for Hydrologic Data

When searching for data in HydroDesktop, you can specify the following filters: region of interest, time period of interest, data source and variables of interest. HydroDesktop then searches the CUAHSI-HIS national catalog of known time series data to find locations of time series that match your search. Locations of time series data that match your search are presented in the map. These results include information that HydroDesktop can use to connect to each individual data provider for data access. You can further filter the results and then choose which data you want to actually download and store in your database.

When you save data to your database, it is stored as a theme. A *theme* is a collection of hydrologic time series data that share a common relationship. A theme can be anything from a geographic space (e.g., Texas, Colorado) to a hydrologic event (e.g., flood, hurricane) to a combination of both (e.g., Texas Flood). Simply put, a theme organizes a collection of related time series. HydroDesktop can save data to a new theme or append data to an existing theme. The workflow for finding data and saving it to a theme is shown in Figure 0-3.
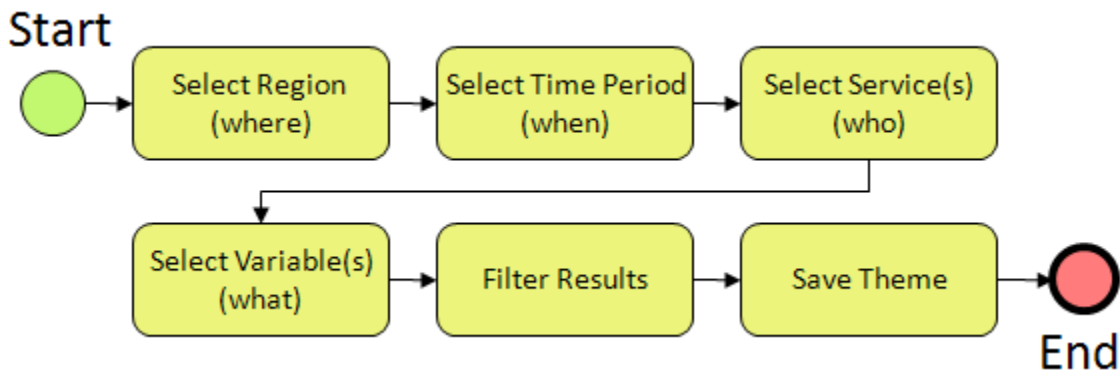
Figure 0-3 Workflow for Searching for Hydrologic Data

In this exercise, you will locate streamflow and temperature data for one water year near Jacob's Well Spring. The county boundary for Hays County is included in the U.S. Counties layer that is already in the map. You'll use this boundary to restrict the area being searched.

1. Click the **Home** tab in the ribbon to activate it.
2. In the **Search Panel** on the right, under the **Area** tab, choose **U.S. Counties** in the list of Active Layers. The map activates this layer while the Search Panel shows the fields in this layer.
3. Under **Select Search Parameter**, scroll through the names until you find Hays, TX. Click **Hays, TX** to select it. The map zooms to the county and highlights it in blue.
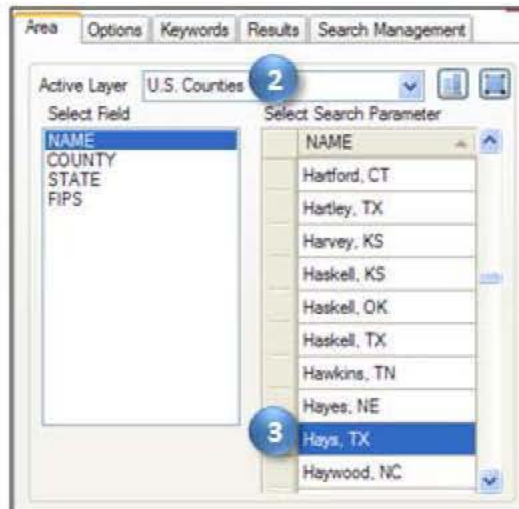
Figure 0-4. Choosing a Search Area

Tip

As you build your query, the current search parameters are shown in the Search Summary at the bottom of the Search Panel. To give yourself more room to work, you can hide the Search Summary by clicking the Hide Search Summary button 🔲. Click the button again to show the Search Summary once it is hidden.

Next you will tell HydroDesktop the date range of time series that you want. For this exercise, search for data available in the 2010 water year (i.e., 10/1/2009 to 9/30/2010).

1. In the **Search Panel**, click the **Options** tab to activate it.
2. Specify a **Start Date** of **10/1/2009** and an **End Date** of **9/30/2010**. You can click and type the numbers in directly, or you can click the drop down arrow next to the date to open an interactive calendar.

Tip

On the Options tab, you can also select specific Web services (i.e., data sources) to query. The default is to search all Web services.

Next you will tell HydroDesktop what hydrologic variables you want. To help you in this regard, HydroDesktop employs a list of official CUAHSI-HIS keywords for hydrologic variables. Data providers use this list when registering with CUAHSI-HIS. This is a lot easier than typing whatever term the data provider may be using internally (e.g., 00060 for USGS streamflow).

3. In the **Search Panel**, click the **Keywords** tab to activate it.
4. Start typing "streamflow" in the **Keywords** text box. The list of keywords automatically selects keywords that match your search.
5. Click **Streamflow** in the list of keywords.
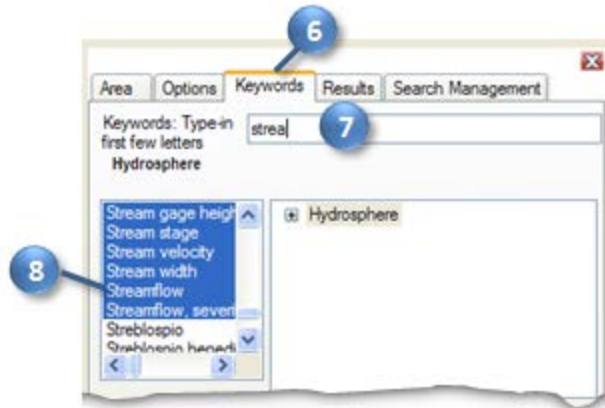
When you click Streamflow, the keyword list automatically jumps to the term "Discharge, stream." It just so happens that "Discharge, stream" is the official keyword for what we think of as streamflow. However, the keyword list also includes synonyms like "streamflow" to make it easier to find the variable you're after. To the right of the keyword list, you can see where stream discharge fits within the overall hierarchy of hydrologic variables.

Now that you've identified the right keyword, you'll tell HydroDesktop to add that keyword to the list of variables for which it will search.

9. In the **Search Panel**, Click the **Add** button ⊕ to add "Discharge, stream" into the Selected Keywords box.

Now you will repeat the process to add the water temperature keyword.

10. In the **Keywords** text box, start typing "temperature."

11. Click **Temperature, water** in the list of keywords.

12. In the **Search Panel**, click the **Add** button to add "Temperature, water" into the Selected Keywords box.

With search parameters set, you will now tell HydroDesktop to run the search for data.

13. In the Search Panel, click **Run Search**.

When you run a search, HydroDesktop asks the CUAHSI-HIS national catalog for descriptions of time series that match your search criteria. At this point, your software is using a remote online resource and bringing back data to display in your map. After HydroDesktop has finished searching for time series, it displays the locations of time series that fit your search in a map layer called "Search Results." Note that there will be several "dots" at a single location in the map if the site represented by a given dot measures more than one time series that matches your search. In other words, each dot in the map represents a time series of data. Different symbols for the dots indicate different data sources.

E-12
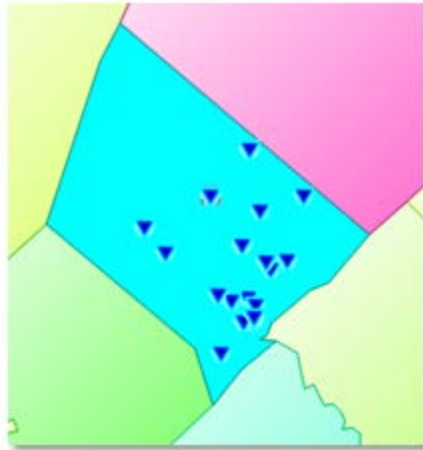
Figure 0-6 Locations of Streamflow and Water Temperature Observations in Hays County, TX

While the search may have seemed fast, remember that your map is only showing where time series of interest are located, and that you haven't actually downloaded any time series values yet. Now you can begin to refine these search results to locate time series that you actually want to download and save to your database.

Selecting Data for Download

To help you identify time series of interest, HydroDesktop includes data and tools to give you a spatial context for the data. One of these is the ability to display online basemaps from ESRI, Bing, and OpenStreetMap. These are beautiful cartographic maps cached at multiple scales which are accessed in real time as you move around in the HydroDesktop map. For this exercise, you will enable the ESRI Hydro Base Map. This map shows a nice blend of hydrologic features and administrative boundaries.

To enable the basemap:
1. On the **Home** tab in the ribbon, find the **Online Basemap** group. Click the drop down list of basemaps and choose **ESRI Hydro Base Map**.
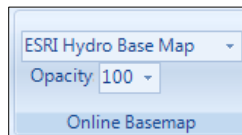
Figure 0-7 Enabling an Online Basemap

In addition to providing spatial context, you can see that this basemap can help you produce a more aesthetically pleasing printed map.

For this exercise, you will work with the sites and variables shown in Table 0-1. You will select the features that represent these time series so that HydroDesktop knows which time series you want to download. These time series are located at or near Jacob's Well Spring.

Table 0-1 Selected Time Series in the San Marcos River Basin

SiteName	VarName	DataType
Blanco Rv nr Kyle, TX	Discharge, cubic feet per second	Average
Jacobs Well Spg nr Wimberley, TX	Discharge, cubic feet per second	Average
Jacobs Well Spg nr Wimberley, TX	Temperature, water, degrees Celsius	Average
Blanco Rv at Halifax Rch nr Kyle, TX	Temperature, water, degrees Celsius	Average

E-13

To select time series for download:

1. In the **Map Contents** on the left, right-click on the **Search Results** layer name and click **View Attributes**.
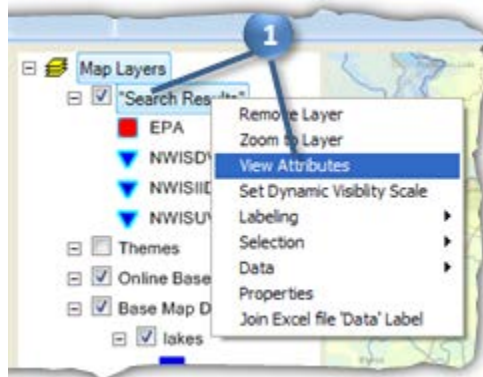
Figure 0-8 Viewing the Search Results Layer Attributes

The Attribute Table Editor opens showing you descriptions of time series in the Search Results layer. You can scroll through the table and resize columns to see the information. Some key columns to note are:

- **SiteName** – The name of the monitoring point where the time series is recorded.
- **VarName** – The name of the variable represented by the time series.
- **DataType** – Some sites report several statistics of data. For example, at Jacob's Well Spring, you can find minimum, maximum, and average streamflow values computed on daily time step.
- **StartDate, EndDate, ValueCount** – These fields give you sense of the overall period of record for a time series and the number of values for its period of record.

As indicated in Table 0-1, you will focus on average streamflow and temperature for Jacob's Well Spring, and you'll also choose a couple of time series from nearby sites for comparison.

2. In the **Attribute Table Editor**, use the values in the **SiteName**, **VarName**, and **DataType** columns to locate rows that match the values in Table 0-1. While holding down the CTRL key, left click on these rows to select them.
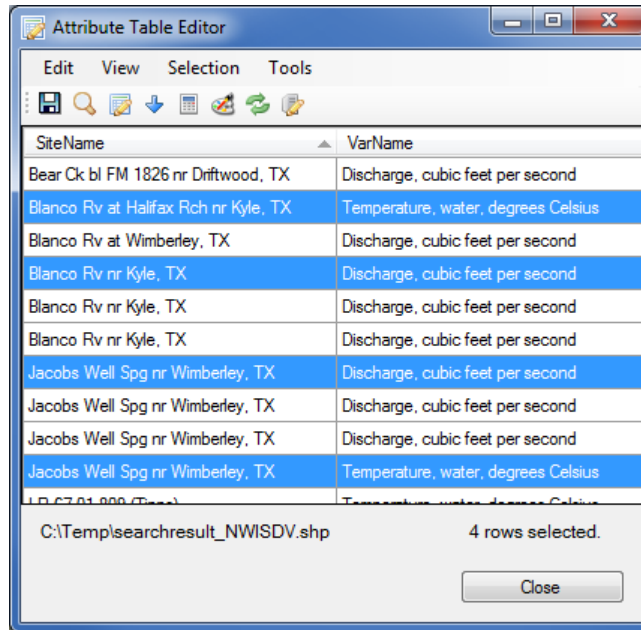
Figure 0-9 Selecting Time Series for Download

    3.   Close the Attribute Table Editor.

With the items selected, you are ready to download the data.

Downloading Data

Recall that when you save data, it is organized into a Theme and stored in the database associated with your HydroDesktop project. You'll now download the data that you selected.

To download data:
    1.   In the **Search Panel**, in the **Results** tab, type "Hays County Data" in the **New Theme** text box, and make sure New Theme is selected.
    2.   Click **Download Data**.
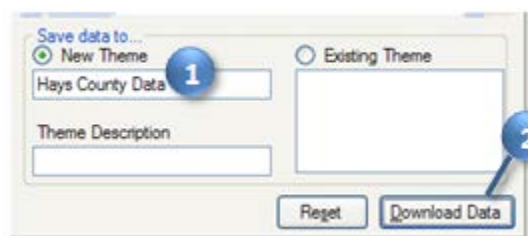
Figure 0-10 Saving a Theme

A download manager opens to show progress of the download.
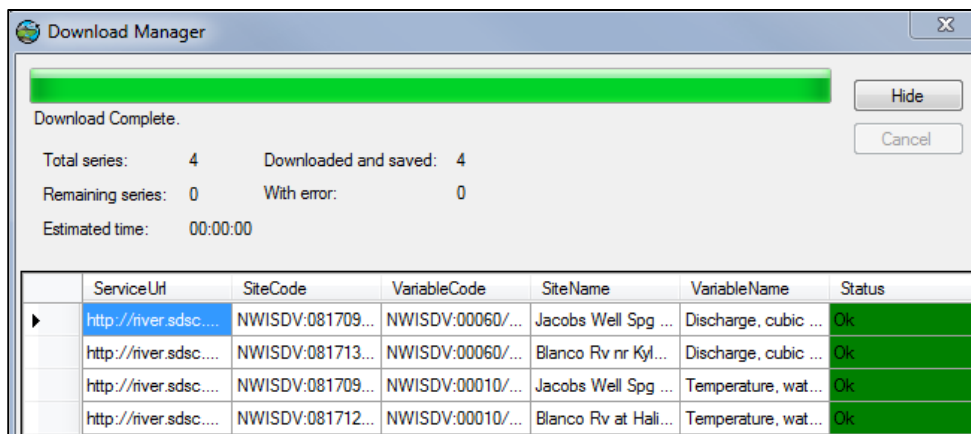
3. Dismiss the message box and hide the download manager when it is finished.

Once the download is complete, the new theme is shown under the Themes group in the Map Contents. If you create additional themes, they will also appear in the Themes group.

Now that you've downloaded the data, you can view the data in both tabular and graph form.

Visualizing Time Series Data

HydroDesktop takes a series-centric view of temporal data, meaning that it provides access to the data at the time series level. An example of a time series is all of the temperature values measured at a certain point on the Blanco River. Let's take a look at the time series that you just downloaded.

To visualize time series data in HydroDesktop:
1. Click the **Graph** tab in the ribbon to activate it.

On the left you will see a list of all the time series in your database. You could use filters to restrict the time series that are shown, but you only retrieved a handful in this exercise so it's fine to leave the default view.
2. Click to place a check next to **Temperature at the Blanco River at Halifax Ranch**.

From the graph you can see how the temperature in the water changes with the seasons throughout the year. Now let's compare this time series with the one for Jacob's Well Spring.
3. Place a check next to **Temperature at Jacob's Well Spring**.

HydroDesktop allows you to visualize multiple time series on the same graph. The plot axes automatically adjust to fit your data. In this example, there is a dramatic difference between the two temperature time series. The one for Jacob's Well Spring shows much less variation throughout the year than the one for the Blanco River at Halifax Ranch.
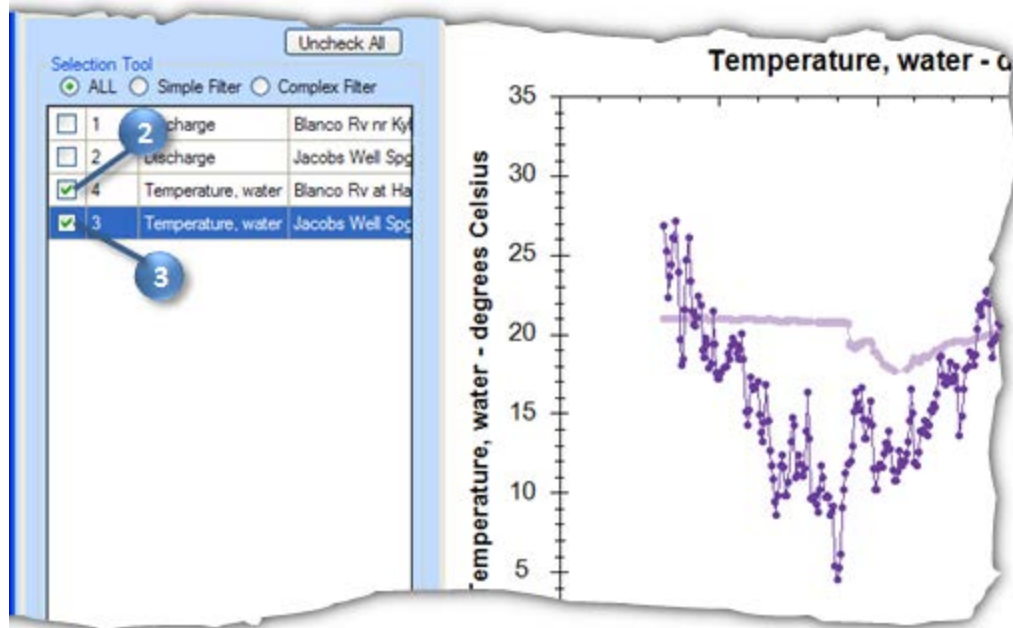
E-16

Figure 0-12 Comparing Temperature Time Series

4. In the ribbon, click **Probability** to show a probability plot of the data, further illustrating the difference between the two time series.
5. Click **Time Series** to restore the time series view.

The dramatic difference in the shapes of the two time series plots is caused by the source of water for these two rivers. The Blanco River at Halifax Ranch is largely a surface water system while Jacob's Well Spring is fueled by groundwater. While the groundwater system does maintain a much more steady temperature than the surface water system, notice how jumps still exist, such as the sudden decrease in temperature in mid-January, 2010. Let's plot flow on this graph to see why this might be happening.

6. Place a check next to **Discharge at Jacob's Well Spring**.

Notice the sharp increase in streamflow around the same time that the water temperature dropped. It seems like the system is experiencing a large influx of surface water, which is colder than the groundwater in the winter. Similarly, you can see an increase in streamflow in late summer, 2010, which results in an increase in water temperature.
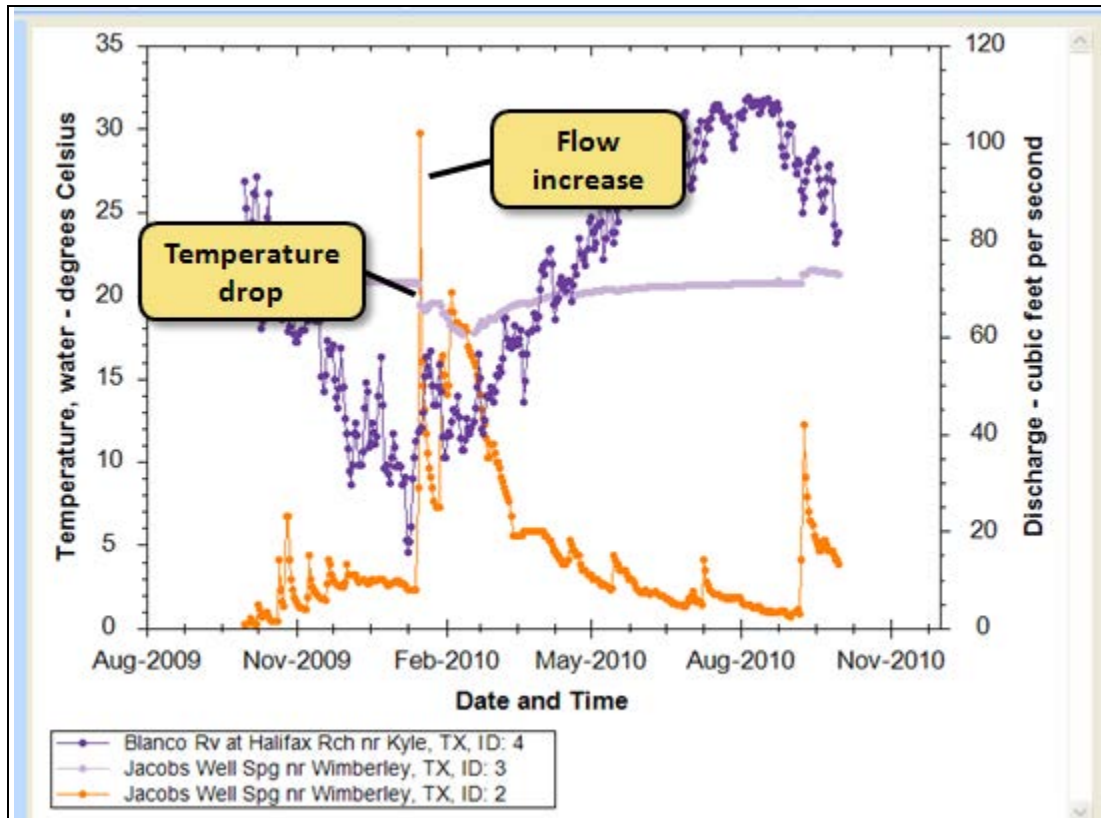
Figure 0-13 Examining Changes in Flow and Temperature

7. Uncheck the two temperature time series, and place a check next to **Discharge at the Blanco River near Kyle**.

The flow of the Blanco River dwarfs that of Jacob's Well Spring, but you can still see increases in flow in the Blanco River at about the same time as those observed in the spring. Also notice the peak flow in September, 2010. This flow is the result of tropical storm Hermine as it swept through Texas.

Labeling Features

Are you wondering where the time series that you just graphed are located? In this portion of the exercise, you will add labels to the map to identify site locations.

1. Click the **Home** tab in the ribbon to activate it.
2. In the Map Layers, uncheck **Search Results** to turn off that layer. Now you should only see the three sites for which you downloaded data.
3. In the Map Layers, right-click on the **Hays County Data** layer name and click **Labeling | Label Setup**.
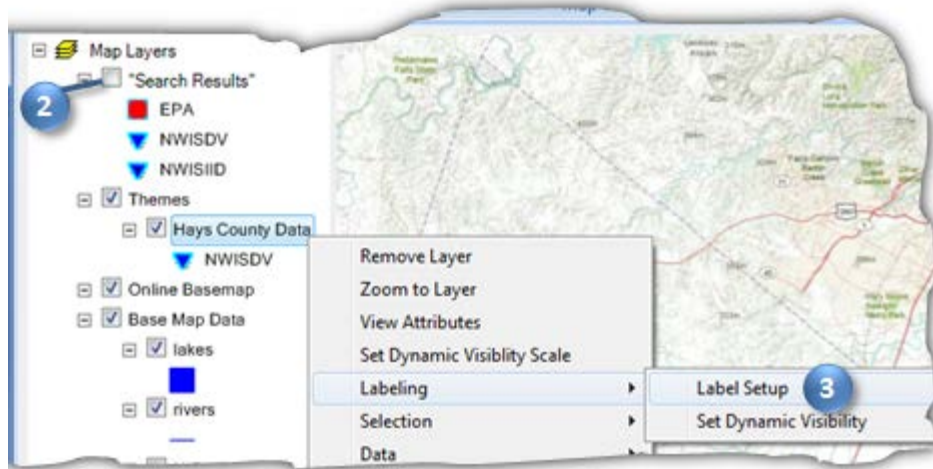
Figure 0-14 Accessing Label Setup

4. In the Feature Labeler dialog, in the list of Field Names, double-click **SiteName** to use that field for the labels. [SiteName] will be added to the text box near the bottom of the dialog.
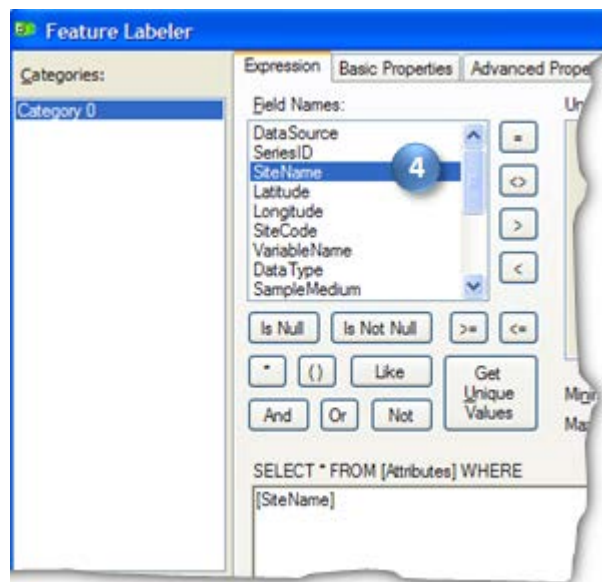
Figure 0-15 Choosing a Field for Labels

5. Click **Apply**, and then click **OK.** Your map should now be labeled according to the settings you have just chosen. You can clearly see Jacob's Well Spring and the two sites along the Blanco River.

E-19

Figure 0-16 Labeling USGS Sites

Delineating Watersheds

At this point, it might be nice to see precipitation data in this watershed for this water year. You can delineate watersheds for any river in the conterminous U.S. using a Web service provided by the EPA. All you have to do is click on the desired watershed outlet location in the map, and then HydroDesktop sends that point location to the EPA service. The service figures out which National Hydrography Dataset (NHD) reach the clicked point is closest to, and then finds all catchments that the reach drains. The catchments are merged into a single watershed and returned to HydroDesktop.

Note that the watershed returned is for the outlet of the entire reach, so if the point you clicked isn't at the reach outlet, then the resulting watershed will include some additional area downstream of your clicked point. Thus, this tool is useful for helping to identify an area of interest but should not be used to determine watershed parameters such as area. Future versions of the tool will support more precise delineation.

In this portion of the exercise you will delineate a watershed for the area draining to Jacob's Well Spring. The watershed delineation tool is part of a HydroDesktop extension called EPA Delineation.

To delineate a watershed for the area draining approximately to the Jacob's Well Spring location:
1. On the **Home** tab, in the **EPA Tool** panel, click **Delineate** to activate the delineation tool.
2. The tool prompts you for where to save the resulting datasets. Accept the defaults by clicking **OK**.
3. Click on the site location for Jacob's Well Spring.

After a moment, the watershed is shown in the map. The NHD reaches flowing to the point that you clicked and the point itself are also shown.

Figure 0-17 Delineated Watershed

If you didn't get the correct watershed delineated then you can activate the tool and try again. It's OK to overwrite previous results.

Note

Recall that the watershed is actually delineated for the outlet of the nearest NHD reach, which happens to be very close to Jacob's Well Spring in this example. Also be aware that the surface watershed you just delineated defines some but not all of the area contributing water to the aquifer for Jacob's Well Spring. However, the area will suffice for this exercise which merely demonstrates how to delineate watersheds and use those watersheds to find data.

With the watershed delineated, now you're ready to search for data in this watershed.

Searching for Additional Data

You can append this data to your current theme by performing another search. In this exercise, you'll search for daily precipitation data in the watershed for Jacob's Well Spring.

To demonstrate how to choose a particular data source for a search, you will select to search for precipitation data from the National Weather Service.

1. Click the **Home** tab in the ribbon to activate it if it is not already active.
2. In the **Search Panel**, under the **Area** tab, choose **watershed** as the active layer.
3. Under **Select Search Parameter**, select the only item present.
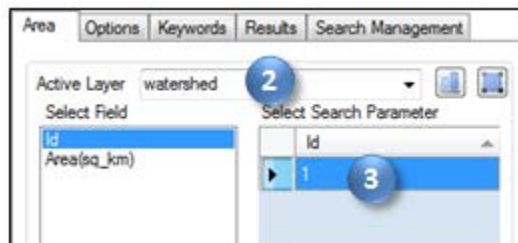
Figure 0-18 Selecting a Watershed to Search

4. In the **Options** tab, click the **Show Web Service Selection Panel** checkbox.
5. Click **Select None**.
6. In the list of services, place a check next to **NWS-WGRFC Daily Multi-sensor Precipitation Estimates**. This service provides precipitation data from the National Weather Service West Gulf River Forecast Center.
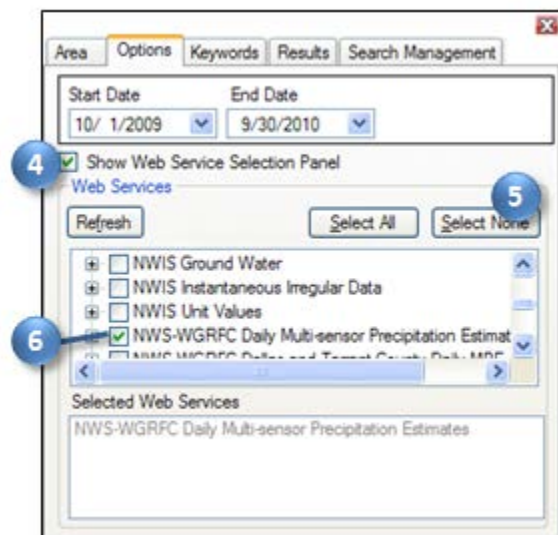
Figure 0-19 Selecting a Single Service to Search

7. In the **Search Panel**, click the **Keywords** tab to activate it.
8. In the Selected Keywords list, select all keywords from the previous search.

9. Click the Remove button ⊖ to remove those keywords from the search.

Figure 0-20 Removing Keywords from a Search

10. In the **Keywords** text box, start typing "precipitation."
11. Click **Precipitation** in the list of keywords.
12. In the **Search Panel**, click the **Add** button to add "Precipitation" into the Selected Keywords box.
13. In the Search Panel, click **Run Search**.

When the search finishes, you'll see some regularly-spaced dots over the watershed. These dots represent the centroids of NEXRAD HRAP cells. In other words, this Web service provides discrete point locations where you can basically sample this gridded rainfall dataset. Each of these points is like a virtual rain gauge. For this exercise, you'll just pick one near Jacob's Well Spring.

Adding Data to a Theme

1. In the ribbon, click the **Select** tool to activate it.
2. In the **Map Layers**, make sure the **Search Results** layer is the only layer selected. The Select tool works with the selected layer in the list of map layers.
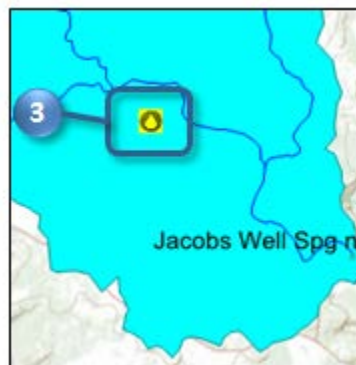3. Draw a box around one or more precipitation sites to select time series for download.

Figure 0-21 Graphically Selecting a Time Series for Download

4. In the **Search Panel**, in the **Results** tab, select **Hays County Data** in the **Existing Theme** text box.
5. Click **Download Data**. Hide the download manager once the download is complete.

E-22

The precipitation data are added to your theme. Now let's view the results.

6. Click the **Graph** tab.
7. Plot a graph with one of the **precipitation** time series you just downloaded and **Discharge from Jacob's Well Spring**. Notice how quickly Jacob's Well Spring responds to rainfall events.
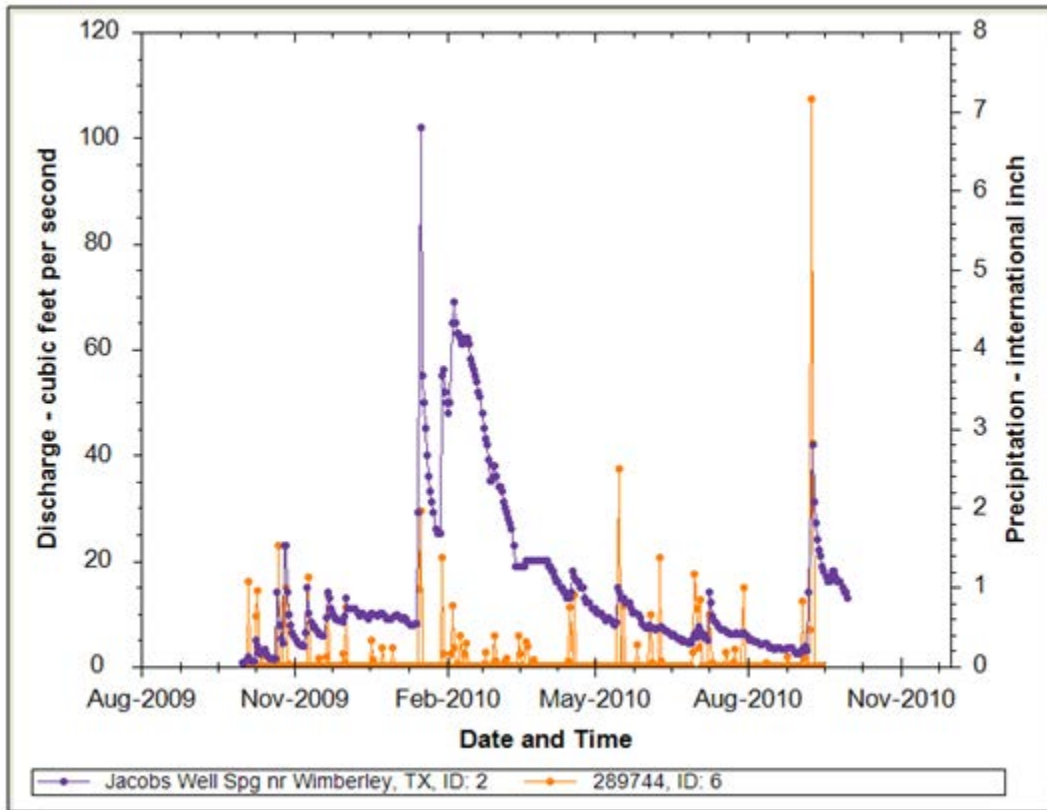

Figure 0-22 Rainfall and Streamflow at Jacob's Well Spring

While HydroDesktop does contain additional analysis capabilities, it can also export data to a text file for use in other programs.

Exporting Data

HydroDesktop can export data to a variety of output file types for further study and analysis. For example, you can export individual time series by placing a check next to them and then right-clicking them in Table or Graph view. For this exercise, you will export all time series for an entire theme.

To export time series data for a theme:

1. Click the **Table** tab in the ribbon to activate it.
2. In the **Data Export** group, click **Export**.
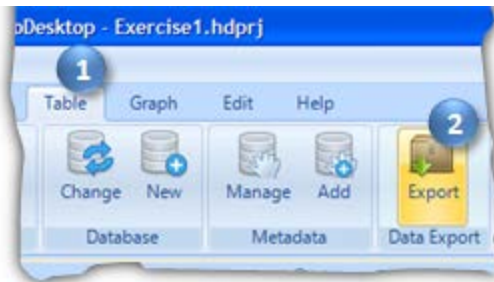
E-23

Figure 0-23 Exporting Data

This tool exports all data in a theme to a delimited text file. In the Export To Text File dialog, notice that your theme name is already selected by default. You can also control the fields that are included in the export and choose a delimiter. For this exercise, you will accept all defaults to produce a comma delimited text file.

3. In the **Export To Text File** dialog, specify the output file location and name.

4. Click **Export Data.**

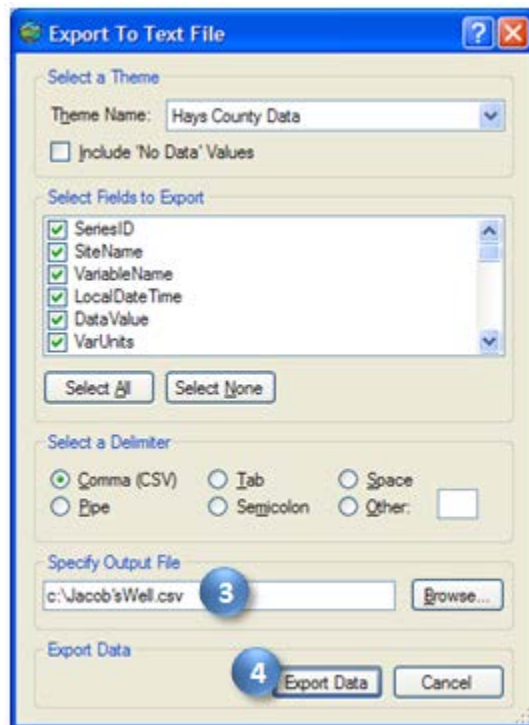5. Close the **Export To Text File** dialog when it is finished.

Figure 0-24 Setting Export Options

6. Find the file on your computer and open it to verify that the data were exported.

Congratulations! With your theme data in hand, you have completed the exercise and learned how to use HydroDesktop to discover and access water data. Feel free to experiment with other functionality such as creating and printing a map, and be sure to give feedback using the Help tab. This concludes the main portion of the exercise. For an example of more advanced analysis, continue with the section below to learn how to use the R statistical environment with HydroDesktop.

Advanced: Analysis with R

The work above has illustrated temperature, precipitation and discharge data and suggested that variations in temperature in Jacob's Well Spring may be related to the mixing of surface and subsurface water sources. In this section, you will use the HydroR plug-in in HydroDesktop to explore this phenomenon. The HydroR plug-in provides an interface between HydroDesktop and the free R statistical software environment.

Enabling the HydroR Plug-in

Once R is installed, you will need to enable the HydroR plug-in in HydroDesktop.

To enable the HydroR plug-in:
1. Click the **HydroDesktop Orb** button.
2. In the **Orb** menu, click **Extensions**. In the extensions list, click **HydroR**. This adds a HydroR tab to the ribbon. You may have to hover the mouse on and off of Extensions to make the list open.
3. Click the **HydroR** tab in the ribbon to activate it.
4. In the **HydroR** tab, click **Start R**.
5. If prompted for the path to R.exe, enter the path where it was installed on your computer. Note that R may include more than one R.exe file. The one you typically need is in the **bin\i386** folder, as in C:\Program Files\R\R-2.13.0\bin\i386\R.exe. Click **OK** once the path is entered. HydroDesktop will remember this path the next time you use the HydroR extension.
6. HydroDesktop needs some additional R libraries. If you are using the HydroR extension for the first time, you may be prompted for a CRAN mirror to download these libraries. Select the mirror closest to your location and click **OK**. The appropriate libraries are downloaded automatically.

R should now start and give you a blank R script in the top panel and R Console in the bottom panel. Standard R commands can be entered in the R Console. HydroR makes it easy to provide R access to the data you have downloaded with HydroDesktop using the buttons on the ribbon in the HydroR tab.
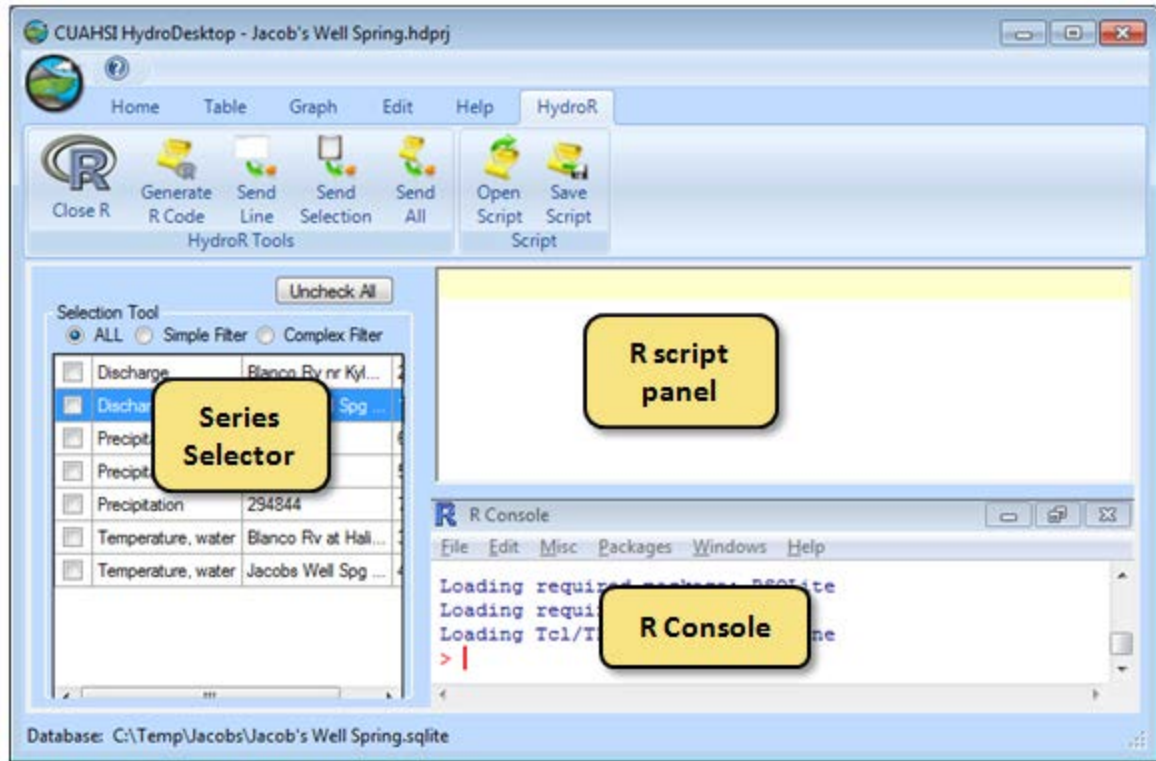
Figure 0-1 HydroR Layout

Plotting a Graph with R

To get familiar with how HydroR works, you'll plot a hydrograph for Jacob's Well Spring.

To plot a graph in HydroR:
1. In the **HydroR** tab, in the list of time series on the left, select a time series that you would like to import as a data frame into R. For this exercise, select **discharge at Jacob's Well Spring**.
2. In the ribbon, click **Generate R code**. The R code to get the selected data series is entered into the script.
3. Click **Send All**. This sends the script text to the console and executes it. The result is an object named **data0**, which contains a list of R data frames.
4. In the **R Console**, enter `labels(data0)` to see a list of data frames that make up data0.

These data frames are basically tables from your HydroDesktop database. The key table we're interested in for this exercise is the DataValues table.
5. To see the first six rows of DataValues, in the **R Console**, enter `head(data0$DataValues)` (R is case sensitive, so type the command exactly as it appears in this text). You may have to scroll up in the R Console to see the full result.

You'll use the LocalDateTime and DataValue columns to provide data for the graph.

E-26

6. To make it easier to access this streamflow time series, in the **R Console**, enter `Q.jacobs=data0$DataValues`. This assigns the DataValues data frame to a variable named **Q.jacobs**.

7. In the **R Console**, enter `plot(Q.jacobs$LocalDateTime,Q.jacobs$DataValue,type="l")` (the type is a lower case L, not a one). A time series plot of the data should appear in an R graphics window, demonstrating that the full capability of R is available to work with the data that has been imported.
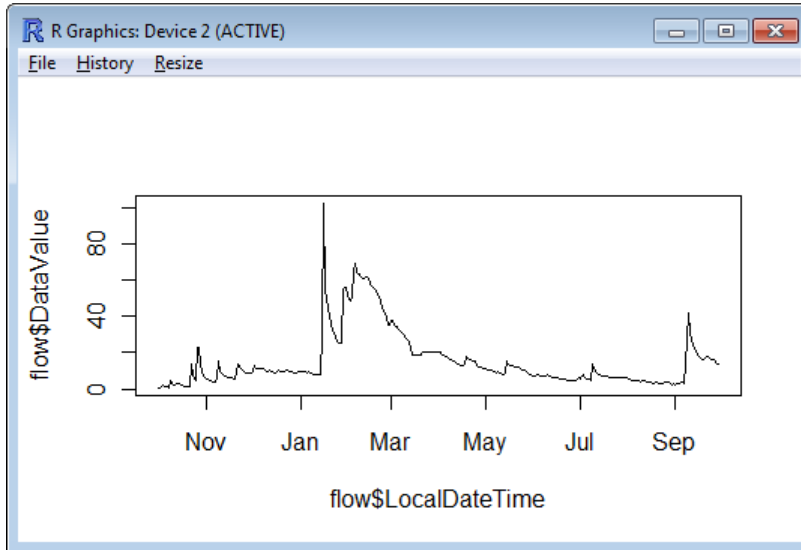

Figure 0-2 Jacob's Well Spring Hydrograph Plotted Using R

Let's add a title to this graph.

8. In the **R Console**, enter `data0$Variable` to see all attributes for the Variable table.

9. In the **R Console**, enter `title(data0$Variable$VariableName)` to add the variable name as the title for the plot.

10. After verifying that the title was added, close the R graphics window showing the graph.

Analyzing Flow in Jacob's Well Spring

Let's use mixing theory to estimate the fractions of Jacob's Well Spring flow that are from the surface and subsurface based on temperature. Assume the surface source has temperature equal to the temperature in Blanco River. Assume groundwater source is at a fixed temperature. The following equations then apply.

Energy Balance: $QT=Q_1T_1+Q_2T_2$
Mass Balance: $Q=Q_1+Q_2$

where Q is discharge in Jacob's Well Spring, T is Temperature in Jacob's Well Spring, T_1 is the temperature of the surface source (assumed equal to Blanco River temperature), and T_2 is the temperature of the subsurface source (assumed constant and taken as the average of last 60 days). Q_1 and Q_2 are the unknown discharge contributions from surface and subsurface sources respectively (Figure 0-3).
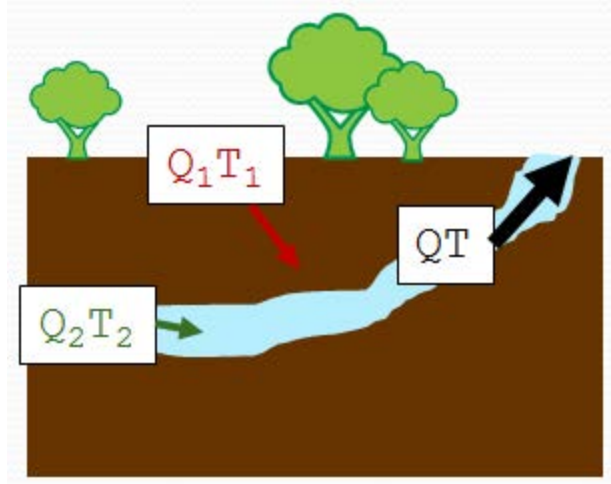
E-27

Figure 0-3 Surface and subsurface contributions to Jacob's Well Spring outflow and temperature

Two linear equations, two unknowns can be easily solved (see your high school algebra book). The solution is

$$Q_1/Q = (T-T_2)/(T_1-T_2)$$

The R scripts in Appendix A use data from HydroDesktop to solve this equation. You'll assign the relevant time series to simple variable names and then use the R scripts to plot a graph representing the amount of flow in Jacob's Well Spring inferred as coming from the surface.

To use the R scripts to compute fractional flow:

1. In the same manner that you created the Q.jacobs variable above and assigned it to be the discharge at Jacob's Well Spring, create and assign the following R variables (remember, the variable names are case sensitive). In other words, for each variable, clear the R script panel, select a series, generate the R code, send it to R, and assign the variable in the R Console.
   a. Q.blanco – Discharge at the Blanco River near Kyle
   b. t.blanco – Water temperature at the Blanco River at Halifax Ranch
   c. t.jacobs – Water temperature at Jacob's Well Spring**

**IMPORTANT

You may encounter a bug when generating R code for water temperature at Jacob's Well Spring. In the R script panel, if the endDate is not "2010-09-30" then edit the script to use "2010-09-30" before sending the script to the R Console.

---

2. Enter **Script 1** found in Appendix A into the **R Console** to execute the script. This script prepares inputs for the analysis and plots graphs of the input temperature and flow.
3. Once you have reviewed the graphs of temperature and streamflow generated by Script 1, close the two R Graphics Windows containing the graphs.

4. Enter **Script 2** found in Appendix A into the **R Console** to execute the script. This script smooths the temperature time series and then performs the analysis to determine the fraction of flow in Jacob's Well Spring from surface water.

The resulting graphs show smoothed temperature time series and the portion of flow in Jacob's Well Spring inferred to be from the surface (the red line in the graph). Note that the analysis requires differences between the assumed groundwater temperature and surface water temperature, so the graph will be missing segments when those temperatures are nearly the same.
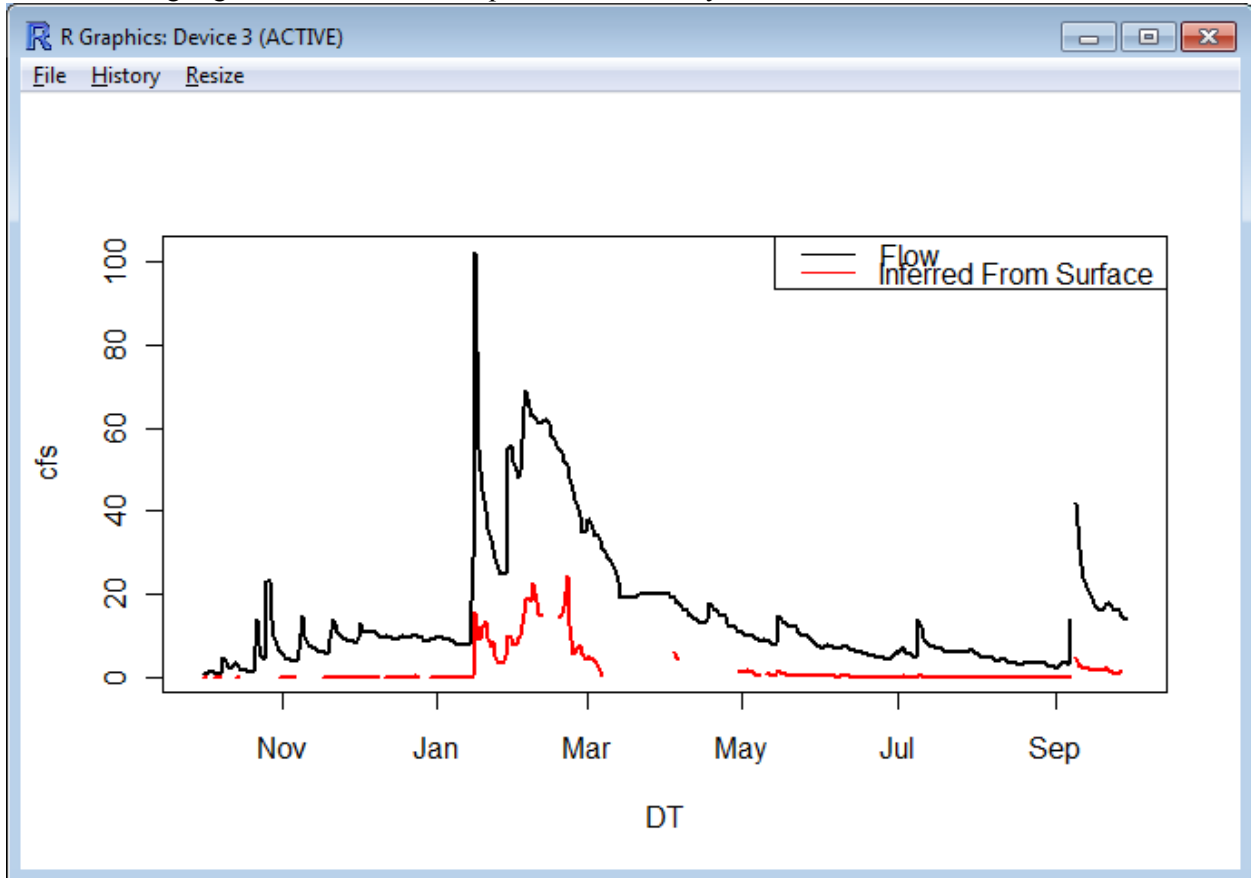

Figure 0-4 Fractional Flow in Jacob's Well Spring

Congratulations! You have completed the exercise and seen how advanced analysis environments such as R can be integrated into HydroDesktop using the power of plug-ins. This concludes the advanced portion of the exercise.

Appendix A: R Scripts

Script 1: Preparing Inputs for Flow Analysis

```
#  SCRIPT 1: PREPARING INPUTS FOR FLOW ANALYSIS

#  This code plots input time series of flow and temperature.
#  The code assumes the following variables have already been set to
#  the DataValues data frame for these time series:
#    Q.jacobs - Discharge at Jacob's Well Spring
#    Q.blanco - Discharge at the Blanco River near Kyle
#    t.jacobs - Water temperature at Jacob's Well Spring
#    t.blanco - Water temperature at the Blanco River at Halifax Ranch
#  The code handles intermittent missing values

#  Start one day earlier because queries seem to be based on UTC
DT = seq(from=as.Date("2009-09-30"),to=as.Date("2010-09-30"),by=1)
ind=match(t.blanco$LocalDateTime,DT)
T1=rep(NA,length(DT))
T1[ind]= t.blanco$DataValue

ind=match(t.jacobs$LocalDateTime,DT)
T=rep(NA,length(DT))
T[ind]=t.jacobs$DataValue
T2 = T[1]  #  The first value

ind=match(Q.jacobs$LocalDateTime,DT)
Q=rep(NA,length(DT))
Q[ind] = Q.jacobs$DataValue
plot(DT,T1,type="l",ylab="T")
lines(DT,T,col=2)
legend("bottomright",c("T Blanco","T Jacobs"),col=c(1,2),lty=1)
windows()
plot(DT,Q,type="l")
```

Script 2: Computing Surface Water Flow Fraction

```
#  SCRIPT 2: COMPUTING SURFACE WATER FLOW FRACTION

#  This script solves the equation Q1/Q = (T-T2)/(T1-T2)
#  and plots a graph showing the portion of flow inferred
#  to be directly from surface water sources in Jacob's
#  Well Spring.
#  Before running this script, you must run SCRIPT 1:
#  PREPARING INPUTS FOR FLOW ANALYSIS

#  Smoothing Blanco River temperature data using lowess
ind=!is.na(T1)  #  array indices of unmissing values
T1l<-lowess(DT[ind],T1[ind],f=0.1)
plot(DT,T1,type="l",ylab="Degrees C")
lines(T1l,col=2)
lines(DT,T,col=3)
```

```r
legend("bottomright",c("Blanco      T","Smoothed     Blanco      T","Jacobs
T"),col=c(1:3),lty=1)
title("Temperatures")

#  Match the dates of the output for use in calculations
ind=match(T1l$x,DT)
T1s=rep(NA,length(DT))
T1s[ind]=T1l$y

#  For results to be reasonable T1 and T2 have to be different.
#  Only evaluate answers when T1 and T2 differ by at least 3 degrees
#  Also only accept positive answers

#  Calculate T2 as average over last 60 days
T2p=c(rep(T[1],59),T)
T2=rep(NA,length(DT))
for(i in 1:length(DT)) T2[i]=mean(T2p[i:(i+59)],na.rm="True")



Q1f=(T-T2)/(T1-T2)  # apply the mixing solution equation
# eliminate answers when temperature difference is less than 3
indna=abs(T1-T2)<3
Q1f[indna]=NA
indna=Q1f< -0.05    # eliminate large negative values
Q1f[indna]=NA

#  Plot the results
windows()
plot(DT,Q,type="l",ylab="cfs",lwd=2)
lines(DT,Q*Q1f,col=2,lwd=2)
legend("topright",c("Flow","Inferred From Surface"),lty=1,col=c(1,2))
title("Jacobs Well Discharge")
```

References

Davidson, S. C. (2008). *Hydrogeological characterization of baseflow to Jacob's Well spring, Hays County, Texas*

    (Master's thesis). Retrieved November 2, 2010, from Hays Trinity Groundwater Conservation District Web

    site: http://haysgroundwater.com/files/Documents/Davidson-08_thesis_Cypress_Crk_Jacobs_Well.pdf.

San Marcos Local News. (2009, March 10). Jacobs Well area to hold incorporation vote. Retrieved November 2,

    2010, from San Marcos Local News Web site: http://www.newstreamz.com/2009/03/10/area-around-

    jacobs-well-to-hold-incorporation-election/.

United States Geological Survey. (2007, March 26). National Surface Water Conference and Hydroacoustics

    Workshop. Retrieved November 1, 2010, from United States Geological Survey Web site:

    http://water.usgs.gov/osw/images/2007_photos/Hydroacoustics.html.