

CUAHSI ODM Tools Version 1.0

Design Specifications

October 27, 2006

Jeffery S. Horsburgh¹

Abstract

The CUAHSI Hydrologic Information System (HIS) Project is developing information technology infrastructure to support hydrologic science. One of the components of the HIS is a point Observations Data Model (ODM), which is a relational database schema that was designed for storing time series data. The purpose of the ODM is to provide a framework for optimizing data storage and retrieval for integrated analysis of information collected by multiple investigators. It is anticipated that the CUAHSI HIS ODM will be implemented by a number of local work groups throughout the country, and that these work groups will use the ODM as a mechanism for publication of individual investigator data and for registering these data with the National HIS. Under this premise, this document provides the design specifications for a set of software tools that will allow administrators of local instances of the ODM to visualize, manage, manipulate, and edit data that have been imported to their local instance of the ODM. The main objective of these tools is to provide administrators of work group instances of the ODM with a set of value added tools that they can use to better manage their data. Additionally, the application and source code developed will provide an example of how applications can be built on top of the CUAHSI HIS ODM.

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) is an organization representing over 100 universities that is supported by the National Science Foundation to develop infrastructure and services for the advancement of hydrologic science in the United States. CUAHSI's mission has several components, one of which is the development of a Hydrologic Information System (HIS) to assemble and synthesize hydrologic data to support hydrologic science development (Maidment, 2005). The CUAHSI 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 (Tarboton et al., 2006). 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. Although designed specifically with hydrologic observation data in mind, this data model has a simple and general structure that it is

¹ Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT, 84322-8200, jeff.horsburgh@usu.edu

hoped will also accommodate a wide range of other data, such as from other environmental observatories or observing networks.

A significant objective of HIS is establishing the cyberinfrastructure foundation, or digital environment required to support experimental watersheds or hydrologic and environmental observatories. The role of the ODM within this objective is to serve as the local repository for point observations data. It is anticipated that the CUAHSI HIS ODM will be implemented by a number of local work groups throughout the country (i.e., at experimental watersheds, test bed project locations, and at hydrologic or environmental observatories), and that these work groups will use the ODM as a mechanism for publication of individual investigator data and for registering these data with the CUAHSI National HIS.

Depending on the level of sophistication of a local workgroup and the expertise that they have available to them, the ODM may be implemented at local workgroup sites by a dedicated Information Technology specialist, or it may be implemented by an individual that knows relatively little about Relational Database Management Systems (RDBMS). In any case, the ODM is simply a database schema consisting of a set of tables, fields, and relationships, and there are currently no data manipulation, analysis, or visualization tools associated with it. Because of this, there is a need for a set of simple tools that will allow ODM administrators, regardless of their level of skill, to manipulate, visualize, and do some simple analysis of data stored within their instance of the ODM. Under these premises, this document provides the design specifications for a set of software tools that will allow administrators of local instances of the ODM to visualize, manage, manipulate, and edit data that have been imported to their local instance of the ODM. The main objective of these tools is to provide administrators of work group instances of the ODM with a set of value added tools that they can use to better manage their data. An additional benefit is that these tools will demonstrate how data visualization and analysis applications can be built on top of the CUAHSI HIS ODM.

CUAHSI ODM Tools Version 1.0

This document describes the design specifications for a software application that will be called ODM Tools Version 1.0, which will subsequently be referred to as “ODM Tools” or as the “ODM Tools application.” This application will be developed in such a way that it is consistent with and compatible with the CUAHSI HIS ODM Version 4.0, which is being released as part of the CUAHSI HIS Version 1.0 package currently under development by the CUAHSI HIS Team. The following sections describe the major features and functionality that will be included in the ODM Tools application.

Features and Functional Requirements

Data Series Query and Export

The CUAHSI HIS ODM Version 4.0 has within it the concept of a “data series.” Each data series in the ODM represents a unique Site, Variable combination, and the SeriesCatalog table in the ODM provides a listing of all of the distinct series of values of a specific variable at a specific site (i.e., observations of water temperature in degrees Celsius measured in the Logan

River near Logan, UT). ODM Tools will include the ability to query an instance of the ODM for specific data series based on information contained in one or more fields in the SeriesCatalog table. Once specific data series are identified, users will then be able to export them to a delimited text file in the CUAHSI HIS MyDB format. Users will be able to preview selected data series in either graphical or tabular format prior to export. The following fields in the ODM SeriesCatalog table will serve as criteria for the Data Series Query and Export tool:

SiteCode: Users will be able to search for data series associated with a particular site by inputting a site code (such as “10109000”) or selecting one from a list of sites populated from data in the ODM.

SiteName: Users will be able to search for data series associated with a particular site by inputting a site name or partial site name (such as “Logan River near State Dam near Logan, UT” or just “Logan River”) or by selecting a site name from a list of sites populated from data in the ODM.

VariableCode: Users will be able to search for data series associated with a particular variable code by inputting a variable code (such as “00060” used to indicate discharge) or by selecting from a list of variables populated from data in the ODM.

VariableName: Users will be able to search for data series associated with a particular variable name by inputting a variable name (such as “Discharge” or “Temperature”) or by selecting from a list of variables populated from data in the ODM.

SampleMedium: Users will be able to search for data series associated with a particular sample medium (such as “Surface Water” or “Ground Water”) by selecting from a list of sample media populated from data in the ODM.

ValueType: Users will be able to search for data series associated with a particular value type (such as “Field Observation” or “Laboratory Measurement”) by selecting from a list of value types populated from data in the ODM.

BeginDateTime and EndDateTime: Users will be able to search for data series that span a particular time period by inputting or selecting a date range.

ValueCount: Users will be able to search for data series that have a minimum or maximum number of observations by inputting criteria for the number of observations.

DataType: Users will be able to search for data series associated with a particular data type (such as “Continuous” or “Instantaneous”) by selecting from a list of data types populated from data in the ODM

GeneralCategory: Users will be able to search for data series associated with a particular general category (such as “Hydrology” or “Water Quality”) by selecting from a list of general categories populated from data in the ODM.

MyDB is a simplified version of the CUAHSI HIS ODM Version 4.0 that consists of a single table with many of the most important fields from the ODM (Table 1). This format was designed by the CUAHSI HIS Team and in the future will serve as the underlying data source for many of the tools developed as part of the HIS Analyst Toolkit under development by the HIS Team. ODM Tools will use the MyDB format for data export.

Table 1. Fields included in the CUAHSI HIS MyDB export file format to be supported by ODM Tools.

MyDB Field	Type	ODM Table	ODM Field	Remarks
ValueID	Number	Values	ValueID	
Value	Number	Values	Value	
DateTime	Date	Values	DateTime	
UTCOffset	Number	Values	UTCOffset	
QualifierDescription	Text	Qualifiers	QualifierDescription	
CensorCode	Text	Values	CensorCode	
OffsetValue	Number	Values	OffsetValue	
OffsetDescription	Text	OffsetTypes	OffsetDescription	
OffsetUnitsName	Text	Units	UnitsName	
Organization	Text	Sources	Organization	
LoadedFrom	Text	<i>No equivalent in ODM</i>	<i>No equivalent in ODM</i>	<i>Specific to MyDB only. Records the procedural call that downloads data to MyDB.</i>
SiteCode	Text	Sites	SiteCode	
SiteName	Text	Sites	SiteName	
Longitude	Number	Sites	Longitude	
Latitude	Number	Sites	Latitude	
LatLonBox	Text	<i>No equivalent in ODM</i>	<i>No equivalent in ODM</i>	<i>No equivalent in ODM. ODM currently only stores point observations.</i>
SRSName	Text	SpatialReference	SRSName	
VariableCode	Text	Variables	VariableCode	
VariableName	Text	Variables	VariableName	
UnitsName	Text	Units	UnitsName	

Metadata Query and Export

ODM Version 4.0 provides the functionality to store a substantial amount of information about each data series contained in the database. This metadata includes information about the site at which the observations were made, information about the variable that was measured and the methods that were used to make the measurements, information about the organization that collected the data, and other supporting information. Much of this information is not included in the MyDB table format. In order to maintain the integrity of data series exported from ODM, ODM Tools will also include functionality to view and export the metadata associated with one or more selected data series.

Users will identify data series using the Data Series Query and Export dialog (see previous sections). Once identified, users will be able to select an option that displays or exports the metadata for a data series. Metadata will be exported to an XML file with a schema based as

closely as possible on the XML outputs from CUAHSI web services. It is anticipated that the metadata XML file will include the following major sections:

General: The general information section will include any metadata information specific to the selected data series that is stored in the ISOMetadata table of the ODM.

Site Information: The site information section will include detailed information about the site at which the selected data series was collected. Information for the site information section will come from the Sites and SpatialReferences tables in ODM.

Variable Information: The variable information section will include detailed information about the variable represented by the selected data series. Information for the variable information section will come from the Variables and Units tables of the ODM.

Source Information: The source information section will include detailed information about the data source associated with the selected data series. Information for the source information section will come from the Sources table of the ODM.

Data Series Plotting and Statistical Summary

ODM Tools will include the ability to visualize selected data series via several different plot types. Users will be able to copy or export the plots generated by ODM Tools as image files that can be inserted into other applications such as Microsoft Word or PowerPoint. The following plot types will be supported:

Time Series: Time series plots provide information about the variability and trends associated with a data series. Options for this plot will include customizing the plotted points and line and user control over the x and y-axis scales.

Histogram: Histograms provide information about the general shape of a data series' distribution.

Probability: Probability plots show the cumulative frequency distribution for a time series dataset and are useful for visualizing quantile and percentile values and for estimating percent exceedance of numeric criterion values (such as numeric water quality standards).

Box and Whisker: Box and whisker plots provide a visual summary of the characteristics of particular groups of values within a selected data series. Options for this plot will include grouping data by monthly, seasonal, annual, or overall (all values) time periods.

A simple statistical summary for selected data series will be presented along with the plots that provides the minimum, maximum, mean, geometric mean, median, 25th and 75th percentile values, number of observations, and number of censored observations. The statistical summary will provide functionality to adjust the descriptive statistics taking into account censored data values for those data series with censored data.

Data Series Editing and Filtering

ODM Tools will include functionality to edit the values and some of the attributes of the values within data series that already exist in the ODM. An example of this is applying a linear drift correction to all or a portion of a data series or filtering data to adjust, flag, or remove erroneous values. Data editing will be based on a visual representation of the data in both plot and tabular formats. Data selected for editing will be highlighted on both the plot and within the table. All data editing will be held in memory until editing for the entire data series is complete, at which point the user will be prompted for whether they want to commit all of the changes to the database as a new data series or replace the original data series. The following editing functions will be supported:

Delete Values: This function will allow users to select values within a data series and permanently delete them from the database. This function will be useful for removing erroneous values from the database, for example deleting values from a water temperature data series for time periods when the temperature sensor was out of the water or buried in sediment.

Adjust Values: This function will allow users to adjust selected values within a data series by adding a constant value or multiplying by a constant value. In addition, this function will allow users to apply a linear drift correction to a data series that adjusts the values within a data series based on a difference between a final value generated by the sensor and an independent observation of the variable being measured (e.g., stage from a pressure transducer compared to a manual reading from a stage plate). This function will be useful to adjust a data series where an incorrect multiplier or offset was used in a datalogger program or where the data series is produced by a sensor that is prone to instrument drift.

Interpolate Values: This function will allow users to replace missing or erroneous actual observations with linearly interpolated values. Interpolation will be based on the value before the first selected value and the value immediately after the last selected value. This function will be useful for filling *brief* data gaps in continuous data series caused by datalogger or sensor malfunctions where interpolation is appropriate.

Flag Values: This function will allow users to assign data qualifying comments to values within a data series. In addition, users will be able to change the value of the QualityControlLevel field associated with the values in the data series to reflect changes in the quality control level of the edited data series. This function will be useful for tagging manually edited values with data qualifying comments.

Filter Values: A filter dialog will be provided that allows users to select values within a data series based on some very simple criteria. Criteria to be supported include: 1) select and highlight values that are greater than or less than a threshold and 2) select and highlight values where the change in value from the previous value is greater than some threshold. This function will be useful for finding and selecting values that may be erroneous and need to be manually adjusted or removed from the database.

Smoothing Values: ODM Tools will allow users to smooth an existing data series already stored within the ODM. Smoothing is useful where the scatter in the data blurs a fundamental underlying pattern. For example, smoothing can be used to provide an estimate when artificial variability has impacted the recorded values. A specific example of this is a water level float in a groundwater well that periodically sticks to the well casing as the water level drops but eventually falls to the water level as gravity overcomes friction. This sticking would cause artificial “steps” in the level data as the float sticks and then suddenly drops, but the overall trend in the data is essentially captured. This functionality will only be available for data series that are continuous and that do not have large data gaps. We will investigate Fourier and LOWESS smoothing algorithms for use in ODM Tools.

Data Series Derivation

ODM Tools will include functionality that will allow users to create entirely new data series (derived data series) from data series that are already stored in an instance of the ODM. The following sections describe the specific data series derivation functionality that will be supported by ODM Tools. Users will have the option to commit new, derived data series to the ODM database with all appropriate metadata information.

Daily Derived Values: ODM Tools will be capable of creating a derived data series by allowing users to generate daily minimum, maximum, and average value statistics from a data series containing more frequent, continuous observations already stored within the ODM. An example of this is computing daily minimum, maximum, and average discharge values for a gage that measures discharge continuously on a 15-minute interval. This function will only be available for continuous data series with consistent time support that is less than 24 hours.

Algebraic Conversions: ODM Tools will support creation of a derived data series by allowing users to perform a simple algebraic calculation on an existing data series already stored within the ODM. An example of this is computing a discharge data series based on a stage data series using a site specific rating curve. A limited number of algebraic equation formats will be supported.

Technical Requirements

The following sections detail specific technical requirements for the ODM Tools application:

Development Environment and Source Code

ODM Tools will be built as a Microsoft Windows application in the Microsoft Visual Studio .Net 2003 development environment. As the ODM Tools application will serve as an example of how applications can be built on top of the CUAHSI HIS ODM, all of the source code for ODM Tools will be commented to enable it to be easily understood by developers unfamiliar with the code. The ODM Tools application and its source code will be made freely available according to the CUAHSI HIS software policy.

Operating System Support

ODM Tools will be tested for use on Microsoft Windows 2000, Windows XP, and Windows 2003 server (32-Bit Version).

Database Support

ODM Tools will be designed to connect to an instance of the CUAHSI HIS ODM Version 4.0 implemented in Microsoft SQL Server 2005 (including SQL Server 2005 Express). The application will provide the user with a simple interface for creating a connection to the database, including server and authentication information. ODM Tools will support connection to either local or remote database servers (i.e., users will be able to install the ODM Tools application on their own PC and connect to either a local or remote server running Microsoft SQL Server 2005).

Third Party Software Components

Rather than recreating specific functionality that can be obtained through free or inexpensive third party software development components, ODM Tools will use existing components where possible. It is anticipated that at least one third party software development component will be used in the design of ODM Tools to provide time series plotting capability. Where possible, freely available or open source components will be used so that developers who wish to edit or recompile the source code for ODM Tools will not have to purchase licenses for any third party software components.

User Interface Requirements

ODM Tools will be a Windows-based application. It will consist of a set of Windows forms that will provide users with a graphical user interface to their instance of the ODM. ODM Tools will be written as a single application, but will be organized such that there are logical and intuitive groupings of functionality. ODM Tools will be menu and button driven, and will not require any user programming or command line functionality.

Installation and Configuration

ODM Tools will be delivered via an executable installation file that can be distributed via compact disk or downloaded from an appropriate website to be determined by CUAHSI. The executable installation will be accompanied by a read me file that will detail any nuances or prerequisites of the installation process and provide users with a resource for troubleshooting the installation. The software installation will install all of the necessary components and files for the ODM Tools application to work, and it will also install a user manual in Adobe PDF format that details the functionality of the ODM Tools. It should be noted that the software installation for ODM Tools will install the software application, but it is left to the user to create an appropriate instance of ODM Version 4.0 within Microsoft SQL Server 2005 for the ODM Tools application to attach to.

Acknowledgements

ODM Tools will use components of the Time Series Analyst application and other source code developed at the Utah State University's Utah Water Research Laboratory by the Environmental Management Research Group. The work of the EMRG programming team is acknowledged and appreciated. In addition, some functionality for ODM Tools has been modeled after specific functionality developed within the United States Geological Survey's Panola Mountain Experimental Watershed 4th Dimension Database. Special thanks are given to USGS and specifically to Brent Aulenbach for providing guidance in the design for ODM Tools.

This material is based upon work supported by the National Science Foundation under Grant No. 03-26064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Tarboton, D.G., J.S. Horsburgh, D.R. Maidment, and B. Jennings. 2006. CUAHIS Community Observations Data Model Working Design Specifications Document – Version 4.
<http://www.cuahsi.org/his/docs/ODM4.pdf>

Maidment, D. R., ed. (2005), Hydrologic Information System Status Report, Version 1, Consortium of Universities for the Advancement of Hydrologic Science, Inc, 224 p,
<http://www.cuahsi.org/docs/HISStatusSept15.pdf>.