CUAHSI ODM Streaming Data Loader
Design Specifications

June 2007

Jeffery S. Horsburgh[1] and David G. Tarboton[2]

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 ODM is to provide a framework for optimizing data storage and retrieval for integrated analysis of information collected by multiple investigators. The CUAHSI HIS ODM is being implemented by a number of local work groups throughout the country, and these work groups are using the ODM as a mechanism for publication of individual investigator data and for registering these data with the National HIS. At many of these sites, investigators are collecting continuous datasets in real time using sensor networks and telemetry systems. This document provides the design specifications for a set of software tools that will allow administrators of local instances of the ODM to automate the loading of these continuous data streams into the ODM. The main objective of the Streaming Data Loader is to provide administrators of work group instances of the ODM with tools that can be used to map their continuous data to the ODM schema and schedule the automated loading of the data into the 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 form disparate sources, both within a single study area or hydrologic observatory

[1] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT, 84322-8200, (435) 797-2946, jeff.horsburgh@usu.edu

[2] Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, (435) 797-3172, david.tarboton@usu.edu

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 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. The CUAHSI HIS ODM is currently being 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 these work groups are using the ODM as a mechanism for publication of individual investigator data and for registering these data with the CUAHSI National HIS. These investigator datasets include those generated by sensor networks and telemetry systems, which in many cases are updated continuously and in real time.

A significant level of time and expertise are required to load data into ODM. This time is compounded for data streams that are updated continuously. Because of this, and because the characteristics of these data streams generally do not change over time (i.e., the observation date/time and value change, but the site, variable, offset, etc. do not change) continuous data streams are ideal for automated data loading. There is a need for a simple set of tools that will allow ODM administrators, regardless of their level of skill, to automate the process of loading these continuous data streams into 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 automate the import of continuous data streams to their local instance of the ODM. These tools will facilitate the mapping of the continuous data to the ODM schema and the scheduling of the data importing so that it is consistent with the data collection frequency.

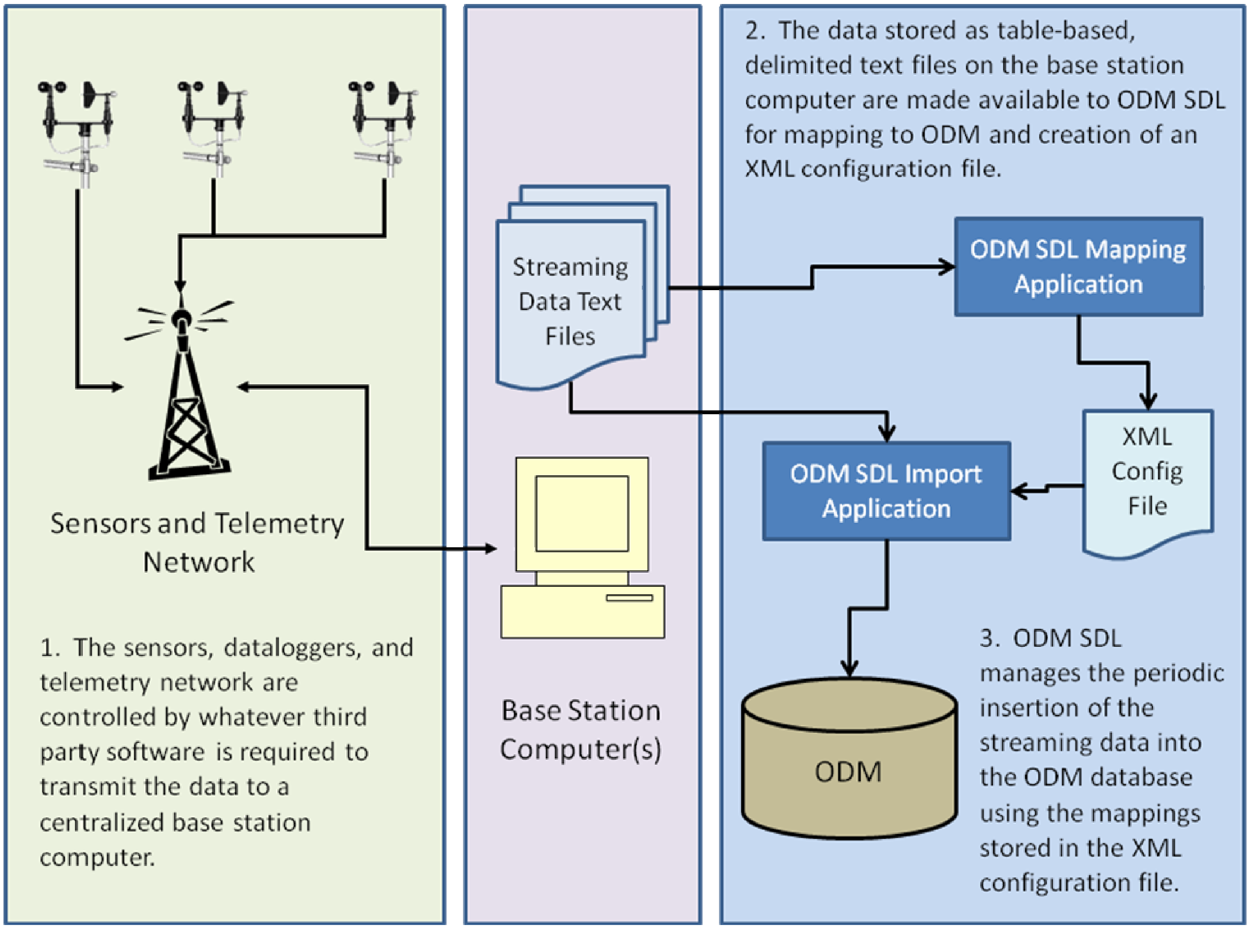
CUAHSI ODM Streaming Data Loader

This document describes the design specifications for a software application that will be called ODM Streaming Data Loader, which will subsequently be referred to as the “ODM SDL.” This application will be developed in such a way that it is consistent with and compatible with the CUAHSI HIS ODM Version 1.0, 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 ODM SDL.

Features and Functional Requirements

It is anticipated that a variety of sensors, datalogger, and telemetry systems will be used within and across the WATERS network of test beds and the planned network of environmental observatories. Because of this, the concept for ODM SDL is that the collection and transmission of the streaming data to a central location will be controlled by whatever proprietary or third party software is required to manage the sensor network. It is also anticipated that the majority of these third party software systems are capable of storing the data once they have been retrieved from remote monitoring sites in a table-based delimited text file.

Once the streaming data have been transmitted to a central location and stored in text files, these text files can then be made available to the ODM SDL for periodic automated upload to an instance of ODM. Figure 1 illustrates this process.



Sensors and Telemetry Network
1. The sensors, dataloggers, and telemetry network are controlled by whatever third party software is required to transmit the data to a centralized base station computer.
Streaming Data Text Files
Base Station Computer(s)
2. The data stored as table-based, delimited text files on the base station computer are made available to ODM SDL for mapping to ODM and creation of an XML configuration file.
ODM SDL Mapping Application
XML Config File
ODM SDL Import Application
ODM
3. ODM SDL manages the periodic insertion of the streaming data into the ODM database using the mappings stored in the XML configuration file.

Figure 1. Schematic of ODM SDL functionality.

ODM SDL will consist of two separate executable applications. The first executable, which will be called the ODM SDL interface executable, will provide the user interface, the interactive data mapping functionality, and all of the general setup functionality. The second executable, which will be called the ODM SDL data import executable, will have no user interface and will be designed to be scheduled for execution by the Windows Task Scheduler. The following sections describe in more detail the specific functionality that is required to automate the process of parsing streaming data into the ODM.

Establishing a Connection to an ODM Database

ODM SDL will provide functionality for connecting to a local or remote ODM database implemented in Microsoft SQL Server 2005. Functionality will be provided for using Windows or SQL Server authentication in the database connection.

Mapping Attributes of Streaming Data Series to Required Fields in ODM

Streaming data produced by a sensor are inherently data series. They are sequential observations of a single variable at a single site using a single method. Because of this, the attributes of a streaming data series can be mapped to ODM once, after which the only thing that remains is to add the data values and value level attributes to the database as they are generated. Generally, most of the data series level attributes required by ODM are not stored in the streaming data file generated by the datalogger and telemetry system. These files typically contain date/time and value pairs with very simple column headers that describe the data in the files. Because the series level attributes cannot be discovered from the data files themselves, an interactive process is needed by which users can create all of the required data series attributes within ODM prior to the streaming data series being loaded.

ODM SDL will use an interactive, wizard based process for mapping the attributes of each streaming data series to required fields in ODM. This user interface will allow users to map streaming data series to sites, variables, methods, etc. that already exist in the ODM database as well as allowing users to create new sites, variables, methods, etc. where they do not already exist in the database. When user interaction with the wizard is complete, the wizard will write all of the required data series level attributes to the ODM database in preparation for loading the data series values. This first step essentially ensures that the ODM database is populated with all of the required series level attributes so that the data values for that data series can be added to the database. This interactive interface will be part of the ODM SDL interface executable.

Creation of the XML Configuration File

The behavior of the automated data loading features of ODM SDL will be controlled by an XML configuration file. This configuration file will contain the connection information for the ODM database and all of the required information about and attributes of each data series to be imported to the ODM database by the ODM SDL. The configuration file will be written at the same time the data series level attributes are being added to the ODM database by the mapping wizard. The configuration file will be created the first time a data series is added for import to an ODM database, and the configuration file will be updated each time a new data series is added for import. The creation and management of the XML configuration file will be part of the ODM SDL interface executable.

The configuration file will essentially store the list of data series to be loaded, information about the location of the streaming data text files on disk or on a network drive, and enough information to create new records in the DataValues table of ODM from the date/time and data value pairs stored in the streaming data text files. This configuration file will enable the complete automation of the actual data loading task (i.e., once the configuration file is created, no user interaction is required to load the streaming data into the ODM database). The ODM SDL data import executable will read the configuration file, find the data files containing data to be loaded, and then merge the data series attributes stored in the configuration file with the dates/times and data values for the data series that are stored in the streaming data text file. The combination of these two will be used to write the new data to the DataValues table of the ODM.

Displaying and Editing the Contents of the XML Configuration File

ODM SDL will provide functionality for displaying or listing the contents of the XML configuration file. This information includes: 1) a list of the mapped streaming data series, 2) information about the text files where the data for these streaming data series are being written by the telemetry system and their location, 3) information about the mapping of each streaming data series to ODM (i.e., site information, variable information, method information, etc.), 4) the schedule for data import for each data series, etc. This functionality will be useful for tracking and managing which data series are being imported to ODM and the mapping information for each.

ODM SDL will also provide functionality for editing the contents of the XML configuration file that controls the automated data import. This is necessary in cases where data series are to be removed from the scheduled data import, or where the characteristics or location of a streaming data text file have changed and need to be updated. The XML configuration file display and editing capabilities will be part of the ODM SDL interface executable.

Displaying the Status of Data Imports for Mapped Data Series

ODM SDL will provide functionality for displaying the status of data import for streaming data series that have been mapped to ODM and that are included in scheduled data import. Status information will include the date and time when the last successful import was run, the number of data values successfully imported to ODM, etc. This functionality will allow data managers to monitor the status and performance of the data importer and troubleshoot any errors that may arise (for example if a datalogger quits reporting new data for upload there is likely something wrong with the monitoring site and it will become apparent when no data are uploaded). The capability to display the status information will be included in the ODM SDL interface executable.

Importing Streaming Data to ODM

ODM SDL will use a separate executable application for performing the actual data value loading operations. The primary reason for this is so that there is no user interaction required to run the actual data value loading functionality. This is critical due to the continuous and unattended nature of streaming data. Once the initial mapping of the streaming data series attributes to ODM is complete, there is no longer a need for any user interaction, and the data value uploads can be run periodically as new observations are made and new data become available. The ODM SDL data import executable will be designed such that it checks the database for the latest data and then only imports data values that are newer than those that are already in the database.

The ODM SDL data import executable will be designed so that it can be scheduled to run on a user defined schedule via the Windows Task Scheduler (i.e., users will be able to schedule the executable to run hourly, daily, or on some other time interval depending on what is most appropriate for the data to be loaded). As a final step in the data series import, the SeriesCatalog table in the ODM will be updated so that it always reflects the most recent data in the database.

Scheduling of Data Series Imports

As stated above, the ODM SDL data import executable will be designed so that it can be scheduled to run periodically using the Windows Task Scheduler. The schedule upon which data imports are run will be user defined such that it can be tailored to the specific needs of the streaming data series to be imported. For example if all of the streaming data series to be updated are on satellite telemetry with data being downloaded only once per day, it would not make sense to schedule the ODM SDL data import executable any more frequently than once per day.

Log File Generation

Because the ODM SDL data import executable will run as a Windows scheduled task with no user interaction, it will be programmed to generate a log file with intuitive messages about the outcome of data loading transactions. This log file will be useful in troubleshooting any problems with the data import. Information contained within the log file will include which data series updates were successful, which ones were unsuccessful and error messages indicating where problems were encountered, numbers of data values added to the ODM database, times at which the uploads were run, etc.

Support for Multiple ODM Databases

The ODM SDL data import executable will assume that an XML configuration file contains instructions for adding a list of streaming data series to a single ODM database. However, the ODM SDL data import executable will be designed such that it can be launched using any properly formatted XML configuration file. This means that if streaming data are to be imported to two separate ODM databases by ODM SDL, two separate XML configuration files will be created and two separate tasks will be set up in the Windows Task Scheduler – one for each ODM database/XML configuration file combination.

Technical Requirements

The following sections detail specific technical requirements for the ODM Streaming Data Loader:

Development Environment and Source Code

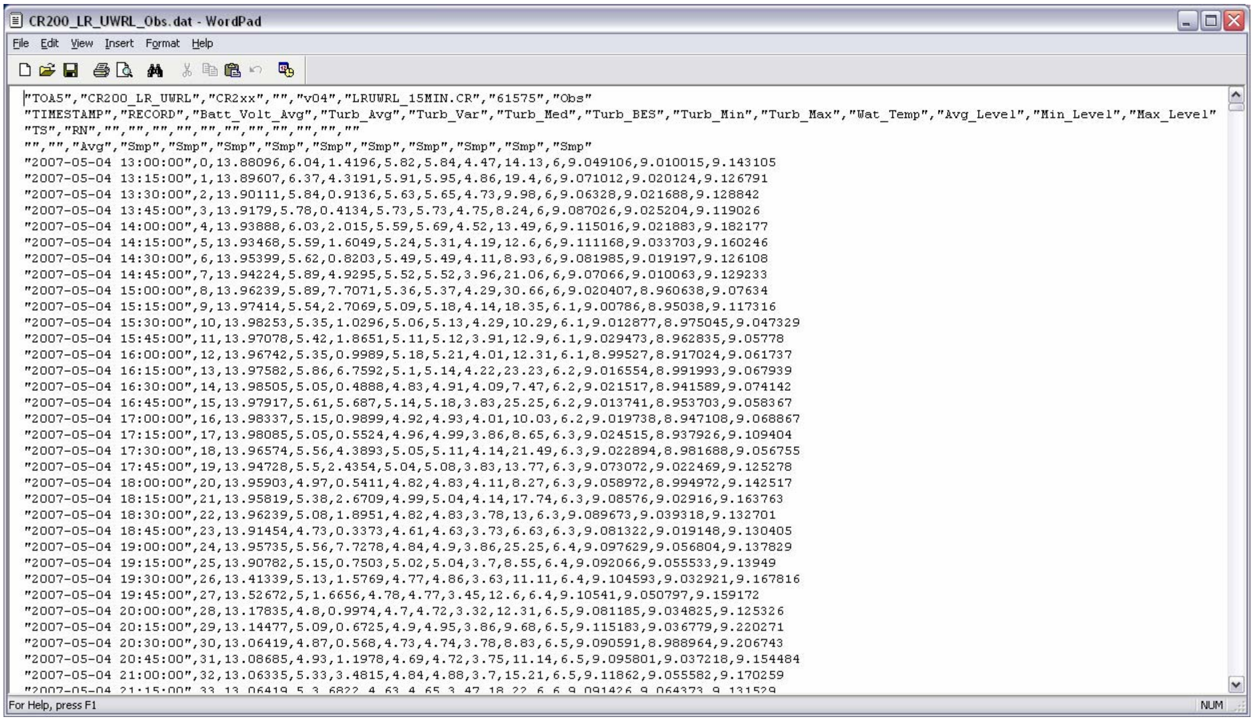
ODM SDL will be built as a Microsoft Windows application in the Microsoft Visual Studio 2005 development environment. The ODM SDL application and its source code will be made freely available according to the CUAHSI HIS software policy.

Operating System Support

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

Support for Streaming Data Files

ODM SDL will support loading data from any table based streaming data file that is formatted as a delimited text file (i.e., table-based data logger files). Multiple data series contained within a single text file will be supported, provided that each data series is contained within a separate column in the file. The following is an example streaming data file with multiple data series in a single data file.


CR200_LR_UWRL_Obs.dat - WordPad
File Edit View Insert Format Help
For Help, press F1
NUM

```
"TOA5","CR200_LR_UWRL","CR2xx","","v04","LRUWRL_15MIN.CR","61575","Obs"
"TIMESTAMP","RECORD","Batt_Volt_Avg","Turb_Avg","Turb_Var","Turb_Med","Turb_BES","Turb_Min","Turb_Max","Wat_Temp","Avg_Level","Min_Level","Max_Level"
"TS","RN","","","","","","","","","","",""
"","","Avg","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp"
"2007-05-04 13:00:00",0,13.88096,6.04,1.4196,5.82,5.84,4.47,14.13,6,9.049106,9.010015,9.143105
"2007-05-04 13:15:00",1,13.89607,6.37,4.3191,5.91,5.95,4.86,19.4,6,9.071012,9.020124,9.126791
"2007-05-04 13:30:00",2,13.90111,5.84,0.9136,5.63,5.65,4.73,9.98,6,9.06328,9.021688,9.128842
"2007-05-04 13:45:00",3,13.9179,5.78,0.4134,5.73,5.73,4.75,8.24,6,9.087026,9.025204,9.119026
"2007-05-04 14:00:00",4,13.93888,6.03,2.015,5.59,5.69,4.52,13.49,6,9.115016,9.021883,9.182177
"2007-05-04 14:15:00",5,13.93468,5.59,1.6049,5.24,5.31,4.19,12.6,6,9.111168,9.033703,9.160246
"2007-05-04 14:30:00",6,13.95399,5.62,0.8203,5.49,5.49,4.11,8.93,6,9.081985,9.019197,9.126108
"2007-05-04 14:45:00",7,13.94224,5.89,4.9295,5.52,5.52,3.96,21.06,6,9.07066,9.010063,9.129233
"2007-05-04 15:00:00",8,13.96239,5.89,7.7071,5.36,5.37,4.29,30.66,6,9.020407,8.960638,9.07634
"2007-05-04 15:15:00",9,13.97414,5.54,2.7069,5.09,5.18,4.14,18.35,6.1,9.00786,8.95038,9.117316
"2007-05-04 15:30:00",10,13.98253,5.35,1.0296,5.06,5.13,4.29,10.29,6.1,9.012877,8.975045,9.047329
"2007-05-04 15:45:00",11,13.97078,5.42,1.8651,5.11,5.12,3.91,12.9,6.1,9.029473,8.962835,9.05778
"2007-05-04 16:00:00",12,13.96742,5.35,0.9989,5.18,5.21,4.01,12.31,6.1,8.99527,8.917024,9.061737
"2007-05-04 16:15:00",13,13.97582,5.86,6.7592,5.1,5.14,4.22,23.23,6.2,9.016554,8.991993,9.067939
"2007-05-04 16:30:00",14,13.98505,5.05,0.4888,4.83,4.91,4.09,7.47,6.2,9.021517,8.941589,9.074142
"2007-05-04 16:45:00",15,13.97917,5.61,5.687,5.14,5.18,3.83,25.25,6.2,9.013741,8.953703,9.058367
"2007-05-04 17:00:00",16,13.98337,5.15,0.9899,4.92,4.93,4.01,10.03,6.2,9.019738,8.947108,9.068867
"2007-05-04 17:15:00",17,13.98085,5.05,0.5524,4.96,4.99,3.86,8.65,6.3,9.024515,8.937926,9.109404
"2007-05-04 17:30:00",18,13.96574,5.56,4.3893,5.05,5.11,4.14,21.49,6.3,9.022894,8.981688,9.056755
"2007-05-04 17:45:00",19,13.94728,5.5,2.4354,5.04,5.08,3.83,13.77,6.3,9.073072,9.022469,9.125278
"2007-05-04 18:00:00",20,13.95903,4.97,0.5411,4.82,4.83,4.11,8.27,6.3,9.058972,8.994972,9.142517
"2007-05-04 18:15:00",21,13.95819,5.38,2.6709,4.99,5.04,4.14,17.74,6.3,9.08576,9.02916,9.163763
"2007-05-04 18:30:00",22,13.96239,5.08,1.8951,4.82,4.83,3.78,13,6.3,9.089673,9.039318,9.132701
"2007-05-04 18:45:00",23,13.91454,4.73,0.3373,4.61,4.63,3.73,6.63,6.3,9.081322,9.019148,9.130405
"2007-05-04 19:00:00",24,13.95735,5.56,7.7278,4.84,4.9,3.86,25.25,6.4,9.097629,9.056804,9.137829
"2007-05-04 19:15:00",25,13.90782,5.15,0.7503,5.02,5.04,3.7,8.55,6.4,9.092066,9.055533,9.13949
"2007-05-04 19:30:00",26,13.41339,5.13,1.5769,4.77,4.86,3.63,11.11,6.4,9.104593,9.032921,9.167816
"2007-05-04 19:45:00",27,13.52672,5,1.6656,4.78,4.77,3.45,12.6,6.4,9.10541,9.050797,9.159172
"2007-05-04 20:00:00",28,13.17835,4.8,0.9974,4.7,4.72,3.32,12.31,6.5,9.081185,9.034825,9.125326
"2007-05-04 20:15:00",29,13.14477,5.09,0.6725,4.9,4.95,3.86,9.68,6.5,9.115183,9.036779,9.220271
"2007-05-04 20:30:00",30,13.06419,4.87,0.568,4.73,4.74,3.78,8.83,6.5,9.090591,8.988964,9.206743
"2007-05-04 20:45:00",31,13.08685,4.93,1.1978,4.69,4.72,3.75,11.14,6.5,9.095801,9.037218,9.154484
"2007-05-04 21:00:00",32,13.06335,5.33,3.4815,4.84,4.88,3.7,15.21,6.5,9.11862,9.055582,9.170259
"2007-05-04 21:15:00",33,13.06419,5,3.6822,4.63,4.65,3.47,18.22,6.6,9.091426,9.064373,9.131529
```

The streaming data files to be processed can be located locally on the same machine as the ODM SDL application, in shared folders on a local area network, or in a web directory accessible via http or ftp. The files and the ODM SDL application will not need to be co-located on the same machine with the ODM SQL Server database.

Database Support

ODM SDL will be designed to connect to an instance of the CUAHSI HIS ODM Version 1.0 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 SDL 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 SDL will use existing components where possible. 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 SDL will not have to purchase licenses for any third party software components.

User Interface Requirements

ODM SDL will be a Microsoft Windows-based application. It will have two separate executables. The first will be for the user interface and will allow users to create a configuration file that maps their data to the ODM schema by interacting with their ODM instance. For example, users will be able to assign their data to an existing site by selecting a site that already exists in their ODM database from a list. This process will take place through a series of windows, menus, and buttons, and will not require any programming by the user. The second executable will perform the actual data loading given the configuration file(s) created by the user interface. The upload executable will not require any user intervention and will not have a user interface. It will be designed such that it can be scheduled to run on a user defined schedule using the Microsoft Windows Task Scheduler.

Installation and Configuration

ODM SDL will be delivered via an executable installation file that can be distributed via compact disk or downloaded from the ODM website at http://water.usu.edu/cuahsi/odm/. The software installation will install all of the necessary components and files for the ODM SDL application to work. It should be noted, however, that the software installation for ODM SDL will install the software application, but it is left to the user to create an appropriate instance of ODM Version 1.0 within Microsoft SQL Server 2005 for the ODM SDL application to attach to.

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.