18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Introducing the Open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop)

Ames, D.P. [1], **J. Horsburgh**[2], **J. Goodall**[3], **T. Whiteaker**[4], **D. Tarboton**[2], **D. Maidment**[4]

[1] Geospatial Software Lab – Center for Advanced Energy Studies, Idaho State Univ., Idaho Falls, Idaho, USA. [2]Utah Water Research Lab, Utah State Univ., Logan, Utah, USA. [3]Dept of Civil and Env. Engr., Univ. of South Carolina, USA. [4]Center for Research in Water Resources, Univ. of Texas at Austin.
Email: amesdani@isu.edu

Abstract: The U.S. National Science Foundation supported Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI) Hydrologic Information System (HIS) project includes extensive development of data storage and delivery tools and standards including WaterML (a language for sharing hydrologic data sets via web services), and HIS Server (a software tool set for delivering WaterML from a server). These and other CUASHI HIS tools have been under development and deployment for several years and together present a relatively complete software "stack", to support the consistent storage and delivery of hydrologic and other environmental observation data. This paper describes the development of a new HIS software tool called "HIS Desktop" and the development of an online open source software development community to update and maintain the software.

HIS Desktop was envisioned as a local (i.e. not server-based) client side software tool that ultimately will run on multiple operating systems and will provide a highly usable level of access to HIS Services. The software will provide several capabilities including data query, map-based visualization, data download, local data maintenance, editing, graphing, data export to selected model-specific data formats, linkage with integrated modeling systems such as OpenMI, and ultimately upload to the HIS server from the local desktop software. As the software is presently in the early stages of development, this paper focuses on design approach and paradigm and is presented to encourage participation in the open development community. Indeed, recognizing the value of community based code development as a means of ensuring end-user adoption, this project has adopted an "iterative" or "spiral" software development approach where 1) the general project requirements and hard boundary conditions are specified at the outset; 2) an initial brief functionality requirements list is developed; 3) the initial limited system is produced primarily by the core funded developer team, but with voluntary external programmer support as it becomes available; 4) testing and bug fixes by the developer team; 5) deployment of an installation package for end-users; 6) collection of bug notices and feature requests from end-users; 7) identification of specific bugs and features to be addressed in a new release; 8) addition of these features by the developer team, etc.

This development approach is the most common approach used by open source projects because of its flexible and dynamic nature. This model is well suited to a community project where it is difficult (and often not useful) to fully-specify the functionality set required for a software release (i.e. as in the "waterfall" development approach), but rather it is desirable to maintain an open structure that can easily be extended through the development of third party plug-ins to support as-yet unknown functions and capabilities, as well as a clear policy on how code is moved into the core system, and how external developers are included in the developer team.

Keywords: *Hydrologic Information System, CUAHSI, observation data, data management*

1. INTRODUCTION

The Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI) is an international organization of universities and hydrologic scientists focused on fostering advancements in the hydrologic sciences, in the broadest sense of that term, by:

- Developing, prioritizing and disseminating a broad-based research and education agenda for the hydrologic sciences derived from a continuous process that engages both research and application professionals;
- Identifying the resources needed to advance this agenda and facilitating the acquisition of these resources for use by the hydrologic sciences community; and
- Enhancing the visibility, appreciation, understanding, and utility of hydrologic science through programs of education, outreach, and technology transfer. (See http://www.cuahsi.org).

CUAHSI's Hydrologic Information Systems (HIS) program has been developed to provide infrastructure to support the interdisciplinary study of hydrologic and related environmental systems, across spatial and temporal scales. The goal of the HIS program is to develop tools that integrate the storage and distribution of data and that facilitate analysis, visualization, and modeling of data.

Within the context of the CUAHSI HIS project, several data storage and delivery tools and standards have been developed as depicted in Figure 1. These include WaterOneFlow web services (a defined set of web services for discovery and download of observation data); WaterML (an XML based language for transmitting observation datasets via web services); HIS Server (a hardware/software system for delivering WaterML from a web server); Observations Data Model (ODM an extensive database schema used to store all data elements associated with point observation data); ODM Tools (a toolkit for managing data in an ODM SQL Server database); HIS Central (a set of web-based tools and central data repository at the San Diego Super Computer Center for hosting and forwarding data from national databases in the WaterML format); HydroExcel (a Microsoft Excel spreadsheet with built-in macros for accessing and retrieving data from a HIS Server into a local file); HydroGet (an extension for ArcGIS that can access and ingest HIS data via a map interface); HydroSeek (a website that aggregates data from multiple HIS Servers and provides ontology based data search capabilities); HydroObjects (programming components used to parse WaterML into standard objects and classes used in custom software projects); and other related tools, services, and data schemas (Goodall et al. 2008, Maidment et al. 2006).

Each of these HIS project components fills a niche needed for specific capabilities related to the cyberinfrastructure surrounding hydrologic information, and together, the tools present a relatively complete

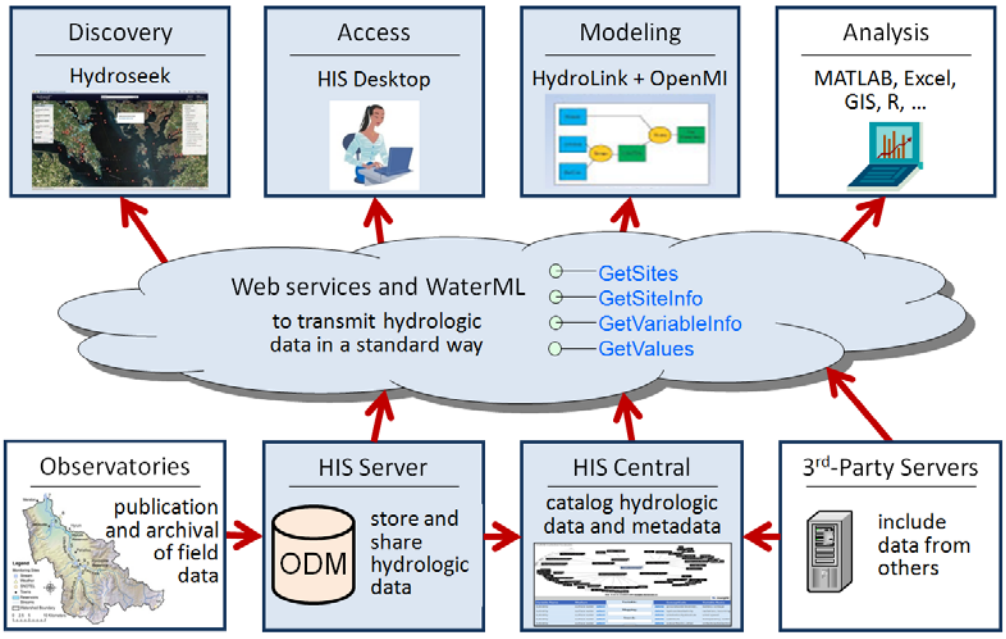

Figure 1. An internet based system to support the sharing of hydrologic data comprising internet connected databases using through web services and software for data discovery, access and publication.

software "stack" to support the consistent storage and delivery of hydrologic and other environmental observation data. This paper presents the development of a new HIS software tool called "HIS Desktop" and the development and support of an online open source software development community to update and maintain the software. HIS Desktop is being developed as a client-side (desktop) software tool that ultimately will run on multiple operating systems and will provide a highly usable level of access to HIS services. The software is envisioned to provide many key capabilities of existing HIS tools (data query, map-based visualization, data download, local data maintenance, editing, graphing, etc.) as well as new capabilities not currently included in any of the existing HIS components (data export to some model-specific data formats, linkage with integrated modeling systems such as OpenMI, and data upload to the HIS server from the local desktop software).

In addition to the core HIS Desktop software design and development effort, a key goal of the HIS Desktop project is the creation of an online community of users and developers who will jointly design, code, bug-test, and deploy the software internationally. This follows an approach common in many open source software development efforts, including the MapWindow GIS project which serves as the GIS component for HIS Desktop (http://www.MapWindow.org). Recognizing the value of community based code development as a means of ensuring end-user adoption, this project is being undertaken using an "iterative" or "spiral" software development approach as shown in Figure 2 where 1) the general project requirements and hard boundary conditions are specified at the outset; 2) an initial brief functionality requirements list is developed; 3) the initial limited system is produced primarily by the core funded developer team, but with voluntary external programmer support as it becomes available; 4) testing and bug fixes by the developer team; 5) deployment of an installation package for end-users; 6) collection of bug notices and feature requests from end-users; 7) identification of specific bugs and features to be addressed in a new release; 8) addition of these features by the developer team, etc.

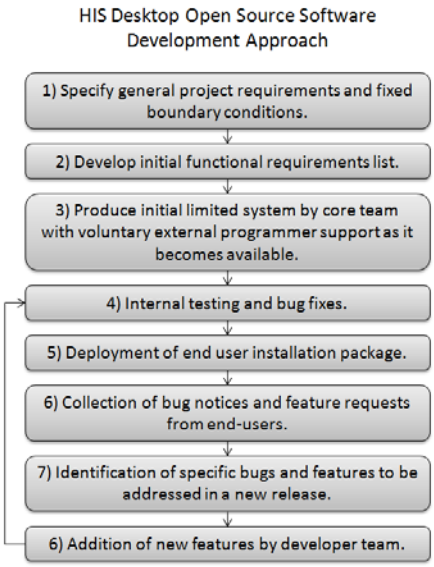

Figure 2. The HIS Desktop open source software development approach including simplified initial specifications followed by iterative addition of functionality.

This development approach is the most common approach used by open source projects because of its flexible and dynamic nature – requirements when the developer team is largely self-selected, self-funded, and self-motivated. Indeed, because of this, it is difficult (and often not useful) to over-specify the functionality set required for a software release (i.e. as is done when following the "waterfall" development approach), but rather it is desirable to maintain an open structure that can easily be extended through the development of third party plug-ins to support as-yet unknown functions and capabilities, as well as a clear policy on how code is moved into the core system, and how external developers are included in the developer team.

2. TARGET USERS AND INTENDED USAGE

HIS Desktop is intended to solve the problem of how to obtain, organize, and manage hydrologic data on a user's computer to support analysis and modeling. This software fills a gap in the HIS project providing a visualization tool for HIS Server based data. The software is intended to be a platform for the integration of HIS data, which can be used in analysis applications such as R, Matlab, and Excel, or in custom code developed by the end user. The HIS Desktop design includes the use of a plug-in architecture and data abstraction layer that will allow extension of the core functionality. Following this approach, the system provides local access to data obtained from distributed data services that are part of the internet-based, service oriented architecture (SOA) that the CUAHSI HIS project has developed for the sharing of HIS data.

It is anticipated that HIS Desktop users will include university faculty, graduate and undergraduate students, K-12 (elementary and high school) students, engineering and scientific consultants, and others. HIS Desktop users may or may not have a technical scientific or computer science background. It is expected that these users will be primarily interested in discovering and retrieving observational data from the HIS system for use in software installed on their local computer. As an open source free software application, HIS Desktop does not require use of specific third party software beyond the operating system.

3. GENERAL SOFTWARE DESIGN

The general design for HIS Desktop and its relationship to other HIS project components are shown in Figure 3. HIS Desktop serves as a common window into observational data published using WaterOneFlow web services. Data discovery is accomplished through searches across a comprehensive metadata catalog maintained at HIS Central and/or individual HIS Servers hosting WaterOneFlow web services. These searches are facilitated by additional web services that expose the metadata catalog and the Hydrologic Ontology maintained at HIS Central. Search results can be further refined to specify datasets that a user would like to download. Data downloads are performed by making GetValues calls (part of the WaterOneFlow web services definition) to the appropriate WaterOneFlow web services. Downloaded data are stored in a desktop data repository database following a relational database schema. This database is accessible to additional tools and software either through an application programmer interface (API) or directly. Visualization and analysis tools that are part of HIS Desktop (e.g. Time Series Analyst) are developed using the API data access method to maintain a level of data access consistency and integrity. Additionally, users can access the data through third party data analysis applications that have the ability to read from a relational database. Such applications include but are not limited to R, MATLAB, and Excel. HIS Desktop includes a number of plug-ins developed by the core project team, and also supports third party plug-ins that follow a standard, well defined plug-in interface described below.

4. KEY HIS DESKTOP FUNCTIONALITY

The primary purpose of HIS Desktop is to facilitate discovery and access of hydrologic data. A secondary purpose is to provide support for data manipulation and synthesis. The user primarily interacts with HIS Desktop via a graphical user interface (GUI) with the functionality described below.

4.1. Data Discovery

HIS Desktop supports two different methods of data discovery: 1) ontology-based discovery across all WaterOneFlow web services that have been registered at HIS Central and for which metadata has been harvested and stored in the HIS Central metadata catalog; and 2) discovery of data within a single WaterOneFlow web service that has not been registered at HIS Central. The first type of data discovery is supported by HIS Central metadata web services that expose the contents of the HIS Central metadata catalog. The second type of data discovery involves making data discovery calls directly to the web service that has not been registered with HIS Central. This approach facilitates both the use of datasets cataloged and documented at HIS Central, as well as use of datasets stored on individual or regional HIS Servers but not necessarily registered with HIS Central.

HIS Central includes a metadata catalog describing the time series datasets served by registered WaterOneFlow web services. This catalog includes the mappings between variables and HIS Ontology concepts. This catalog is automatically updated weekly and represents a comprehensive listing of data published using WaterOneFlow services and registered at HIS Central. The contents of the HIS Central metadata catalog are exposed by a web service API that is currently under development. At a minimum, this metadata catalog web services API

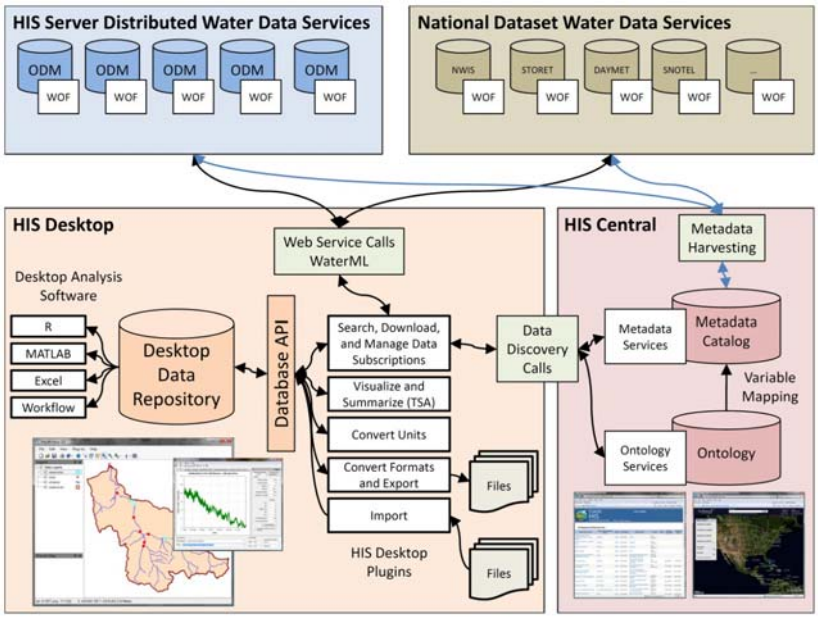

Figure 1. HIS Desktop design and relationship to other HIS components.

will provide methods for retrieving the following information:

- The full metadata description (including the WSDL URL) for all WaterOneFlow web services registered at HIS Central.
- A listing of all searchable keywords/concepts from the HIS Ontology.
- The full metadata description for all data that meet certain spatial, temporal, and variable search criteria.

HIS Desktop uses the methods from the HIS Central metadata catalog API to provide search capability across the metadata catalog to determine relevant data series for a specific user. HIS Desktop presents users with a data discovery form that enables them to input the following search criteria. All of these criteria are optional, but at least one must be specified.

- A latitude/longitude bounding box to serve as the spatial constraint on the query. The box can be input by typing in coordinates, by drawing a rectangle on the HIS Desktop map, or by selecting a polygon feature from one of the layers in the HIS Desktop map (e.g., a watershed boundary – the extent of the feature would be converted to a latitude/longitude box).
- A searchable concept from the HIS Ontology (to be input by the user or selected from a list)
- A begin date and end date to serve as the temporal constraint on the query.
- A minimum number of observations (only data series that have more than this minimum number for the entire data record will be selected, regardless of time window specified.)
- A list of WaterOneFlow web services to include in the search. This will be a user-specified subset of the web service registered at HIS Central that constrains search results to only a selected set of web services.

The result of a data discovery query using the HIS Central metadata catalog is the full metadata description for a listing of all of the data series cataloged at HIS Central that meet the search criteria. For example, a user may choose to search all of Idaho for streamflow data. The results of the search will be a list of sites and data series that meet the criteria. This user is given the opportunity to subset the results to the data series of particular interest, i.e. after seeing a map of the locations of several hundred streamflow gauge sites in Idaho, the user may choose to only retrieve data for sites that meet some additional condition. Following a software "wizard" tool, the user is then be guided to the download button which actually retrieves the selected datasets from their respective WaterOneFlow services and caches the data locally. Users are given the opportunity to "flag" or "label" the datasets that are downloaded to make them easier to find and work with in the future. In addition, the user is asked to put the data locations into a thematic data set on the local machine for layered GIS viewing and interaction.

Figure 4 shows the "simple search" HIS Desktop plug-in that performs a data query against HIS Data using only parameter/variable and area of interest based searching.

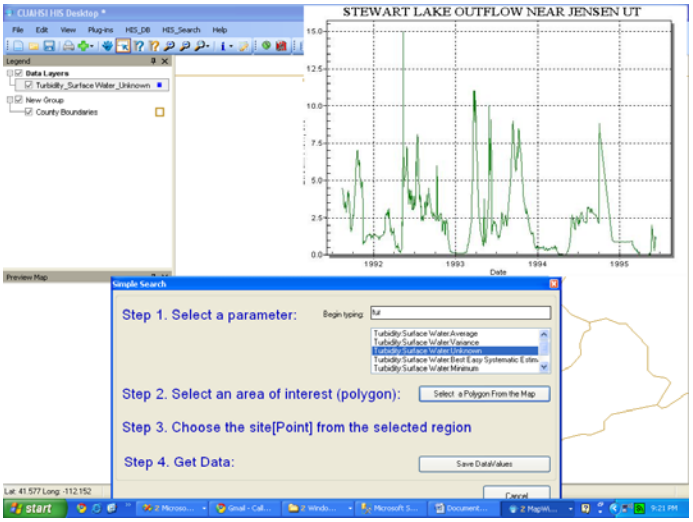

Figure 4. HIS Desktop prototype with "Simple Search" plug-in and time series viewer plug-in.

4.2. Data Download

The goal of the HIS Desktop data download functionality is to retrieve observational data series that have been identified for download using the data discovery tools described above and to create a local cache copy of the data in the desktop data database. Through the underlying MapWindow GIS components (version 6), HIS Desktop can connect to, download and display GIS datasets published using OGC Web Feature Services (WFS), Web Coverage Services (WCS), and Web Map Services (WMS) (Michaelis and Ames 2008).

The result of data discovery is a set of metadata describing data series that have been identified by a user for download. Using this list, HIS Desktop issues GetValues calls to retrieve each data series in WaterML format. As a user selected option, HIS Desktop saves a copy of the result of each GetValues call as a WaterML formatted XML file on the user's hard drive. Next, HIS Desktop parses each of the WaterML results into the HIS Desktop data repository database. The purpose of saving the WaterML files is to preserve the data as they were retrieved from the web service when the GetValues call was made as part of data provenance. The purpose of loading the data into the data repository database is to facilitate and enable analysis and manipulation of the data.

The data repository database has a relational structure and is implemented within a relational database management system (RDBMS), serving as a local cache copy of the data that have been retrieved. The relational schema of the data repository database is semantically similar to the CUAHSI ODM database design (Horsburgh et al. 2008) with similar naming conventions and data types, but has been modified and extended to facilitate management of the data series that have been downloaded and storage of provenance information. Figure 5 shows the relational schema of the HIS Desktop data repository database.

The data repository database is capable of storing all of the information encoded within WaterML files resulting from GetValues calls and also supports the storage of provenance information that includes the following list. HIS Desktop stores and manages this provenance information within the data repository database.

- Where was the data obtained, i.e., which web service?
- How was the data obtained, e.g., from a web service or from a local data import?
- The query that resulted in the data that was loaded (the GetValues call used to get the data)
- A pointer to the WaterML file from which the data originated (the file will be cached locally)
- The date on which the data were loaded
- The last date on which the data were checked for updates
- The last date on which the data were updated with new data
- What has been done to the data since it was added to the database

4.3. Data Visualization, Manipulation, and Export

HIS Desktop supports visualization of both geospatial and time series data. Geospatial data visualization is enabled through an interactive GIS map using the open source MapWindow GIS components (Ames et al. 2008) and 3rd party MapWindow plug-ins. Visualization of observational data is provided through a variety of plots using the open source Zed Graph plotting package and is focused on exploratory data analysis for data series that are downloaded and stored in the HIS Desktop data repository. The HIS Desktop interactive map is used for displaying and manipulating spatial datasets as well as for setting the context for data discovery. As described in the sections above, an area of interest is often used as a spatial filter for narrowing a search for data. Therefore, the HIS Desktop interactive map enables the user to set the geographic context for data discovery and access by enabling users to draw a bounding box or select a polygon feature from one of the GIS layers in the map (e.g., state boundaries, watershed boundaries, etc.) within which they would like to conduct their search. HIS Desktop uses functions in the underlying MapWindow GIS system to provide users with the ability to visualize and manipulate spatial datasets. MapWindow supports a variety of vector, raster, and image GIS data types, and includes functionality for navigating the map as well as many other GIS tools and features. All of the functionality provided by MapWindow for the visualization and manipulation of GIS datasets will be available within HIS Desktop.

Once time series of observational data have been retrieved and stored in the desktop data repository database, HIS Desktop provides users with tools for visualizing and analyzing the data. HIS Desktop maintains a GIS data layer showing the locations of the sites for which data have been downloaded to the desktop data repository database. This layer is dynamically built from the data repository database each time data are downloaded. Users are able to select a site on the interactive map and launch the time series visualization and analysis tool with data populated for the selected site. A variety of plot types are available for visualizing time series data at a selected site. These include time series, histogram, box-and-whisker, and probability plots for a selected time series. The HIS Desktop time series visualization and analysis tool also enables users to view a selected time series in a simple tabular view as well as calculating simple descriptive statistics (minimum, maximum, mean, percentile values, etc.) for the selected time series.

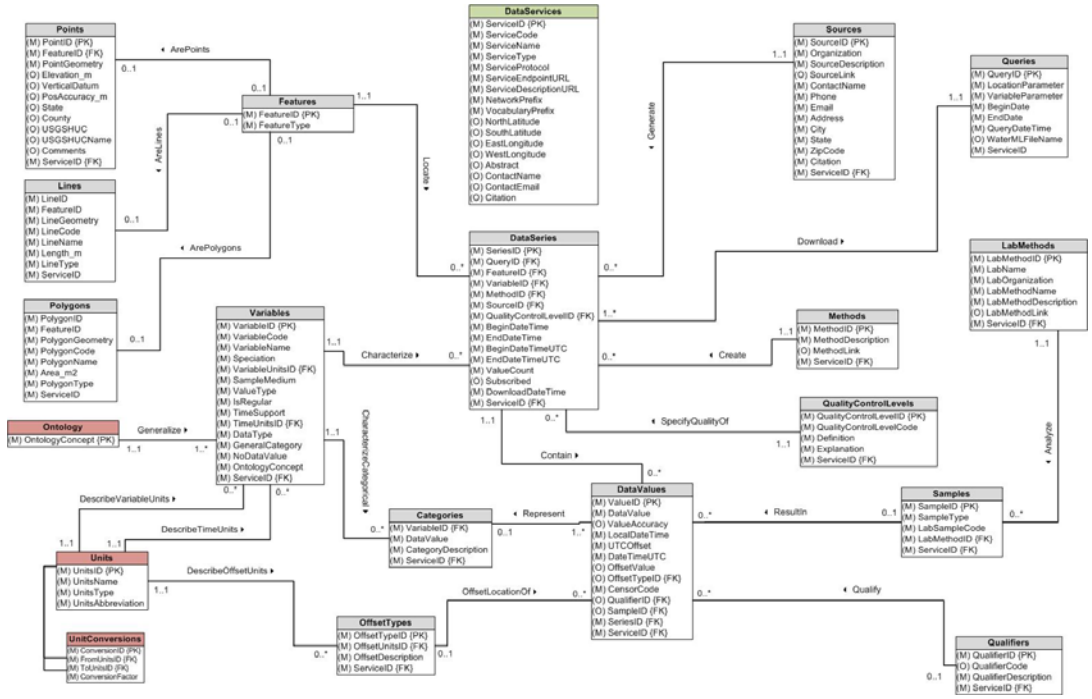

Figure 5. HIS Desktop data repository relational database design.

HIS Desktop includes a data transformation plug-in for creating new data series from existing data series stored in the data repository database. Derivation of new data series will include conversion of units and data aggregation (e.g., converting hourly data to daily data). Additionally, a data export plug-in allows users to export selected observation data from the local database to a file in comma separated file format.

5. OPEN SOURCE COMMUNITY DEVELOPMENT

The CUAHSI HIS Desktop project is under active development and a community of users/developers working on the project is growing at http://his.cuahsi.org/. Here project participants, both from the core NSF funded team and volunteers from the hydrologic sciences community share a discussion forum, bug tracing system, documentation WIKI, and an open Subversion code sharing repository. Any interested parties are invited to visit the project web site, download the source code and join in the development and testing activities related to this project. It is expected that the simple plug-in architecture will encourage and facilitate third party development of plug-ins that significantly extend the base HIS Desktop application, making full use of all of the data retrieval and storage mechanisms in the initial version of HIS Desktop.

REFERENCES

Ames, D.P., Michaelis, C., Anselmo, A., Chen, L., and Dunsford, H., (2008), MapWindow GIS. *Encyclopedia of GIS*. Sashi Shekhar and Hui Xiong (Editors). Springer, New York, pp. 633-634.

Goodall, J. L., J. S. Horsburgh, T. L. Whiteaker, D. R. Maidment and I. Zaslavsky, (2008), A first approach to web services for the National Water Information System, *Env. Modelling & Software*, 23(4): 404

Horsburgh, J. S., D. G. Tarboton, D. R. Maidment and I. Zaslavsky, (2008), A Relational Model for Environmental and Water Resources Data, *Water Resour .Res*., 44:W05406.

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 Modelling and Software*, 24, 879-888, doi:10.1016/j.envsoft.2009.01.002.

Maidment, D. R., Zaslavsky, I. and J. S. Horsburgh, (2006), Hydrologic Data Access Using Web Services, *Southwest Hydrology*, 5(3).

Michaelis, C., and Ames, D.P., (2008), Web Mapping Service (WMS) Web Feature Service (WFS) Web Processing Service (WPS). In: *Encyclopedia of GIS*. Sashi Shekhar and Hui Xiong (Editors). Springer, New York, pp. 1259-1261.