CUAHSI

UNIVERSITIES ALLIED FOR WATER RESEARCH

HydroServer For Azure

Cloud Publication
with the
CUAHSI Water Data Center

Version 1.0

Prepared by Jon Pollak, CUAHSI User Support Specialist

June 2014

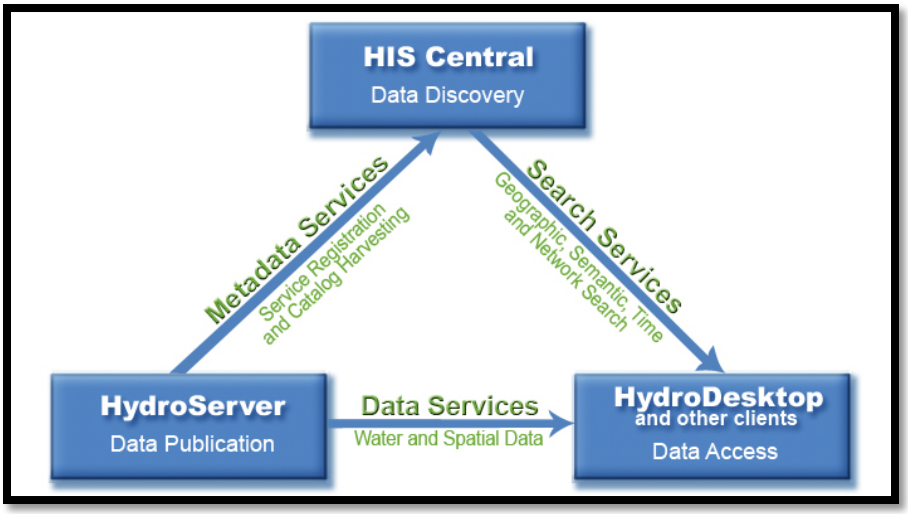
CONTENTS

DISCLAIMER

The CUAHSI Water Data Center (WDC) is a facility funded by the National Science Foundation to enable access to water data holdings, create and maintain data discovery tools, and provide archival and data publishing capabilities. The WDC is funded by National Science Foundation grant number 1248152. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction and Purpose of This Document

The CUAHSI Hydrologic Information System is has been developed in order to enable the sharing of water data between researchers and groups located anywhere in the world. It is composed of three parts. The first component is a data server (HydroServer), which holds and provides data upon request. The second is the metadata catalog (HIS Central), which facilitates search using metadata. The last component is the client used to search and retrieve data (e.g. HydroDesktop).



Above: The three components and linkages of the CUAHSI Hydrologic Information System.

Web services, or computer programs that request and transmit data, facilitate the interaction between these three components.

The focus of this document is on the data storage component of the CUAHSI HIS, HydroServer, which has been deployed in the Microsoft Azure Cloud. The content of this document describes how to request using these cloud services from CUAHSI, how to populate a database in the cloud, and how to register as a data source on the CUAHSI Water Data Center's central metadata catalog, HIS Central.

The interface for CUAHSI's Azure-based HydroServer is located on the domain http://hydroportal.cuahsi.org.

Workflow for Publishing Data in CUAHSI HIS With Azure Software and Services

There are four general steps for publishing data in Azure with the CUAHSI WDC:

1. Requesting a Database
2. Review by CUAHSI WDC Staff

3. Population of the Database
4. Registering the Database as a Data Source on HIS Central

USER REQUEST

Access to the CUAHSI WDC's hosting services is limited to approved data publishers. Requests for databases can be sent to help@cuahsi.org and must include the following information:

- Name of person requesting the database
- Affiliation of person requesting the database
- Contact email
- Google account that will be used for authentication in the WDC system[1]
- Proposed name of network of sites (A unique code associated with the data source)

APPROVAL AND NOTIFICATION BY CUAHSI WDC STAFF

Once a request has been approved, CUAHSI WDC staff will take the following actions:

1. Deploy an instance of the ODM Database in Azure
2. Deploy an instance of WaterOneFlow web services in Azure
3. Configure the ODM database and WaterOneFlow web service so that they are linked to the uploading application and authenticated through the user's Google account
4. Notify the user when the above items have been successfully provisioned and provide user with URL for data service

POPULATING THE DATABASE

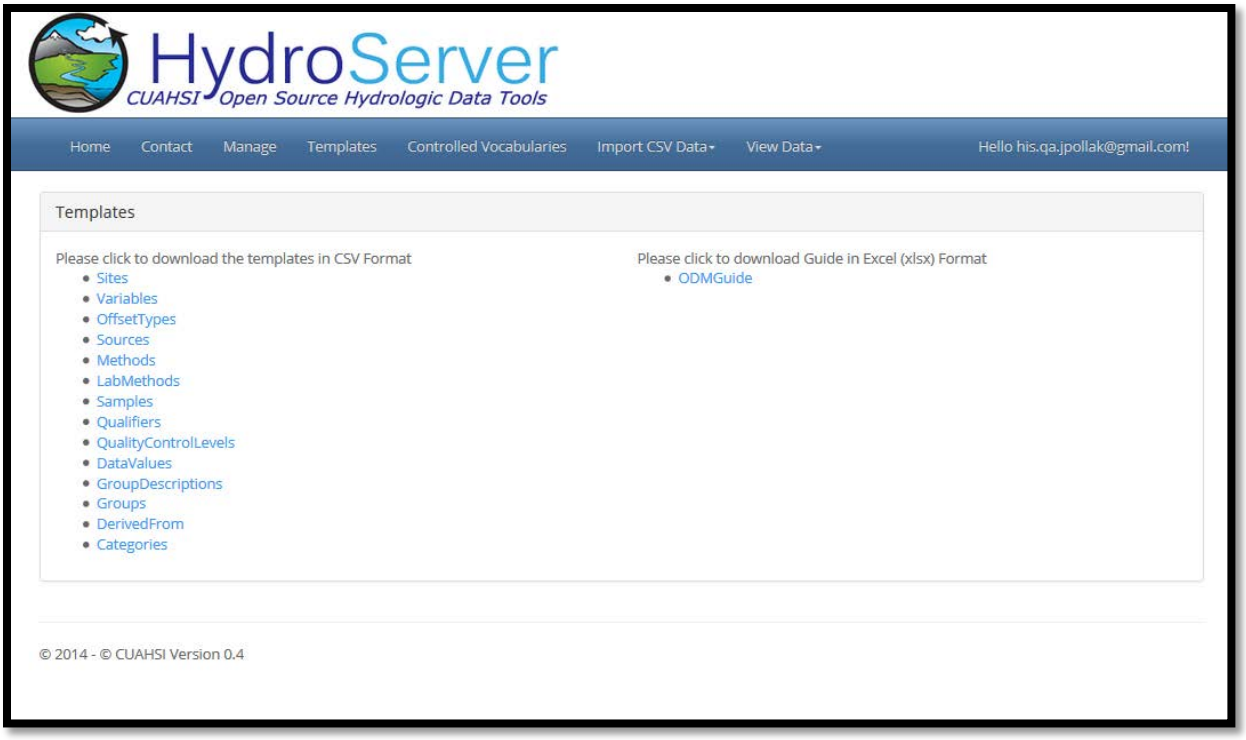
Populating the database occurs in two steps:

1. Formatting data to comply with the Observations Data Model (ODM) using CUAHSI provided templates
2. Uploading templates that contain metadata and data

FORMATTING DATA

The ODM Uploader for Azure contains templates in the form of .csv files under the Templates tab of the website, which are seen in the screenshot below.

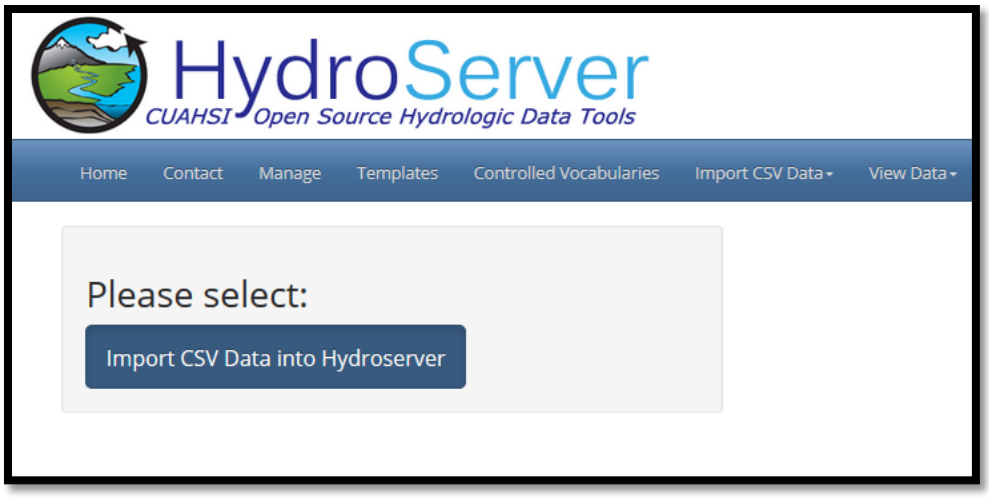
[1] This can either be a .Gmail domain or another domain that is using Google's authentication protocol (such as an institution that uses Google Apps as its email provider).



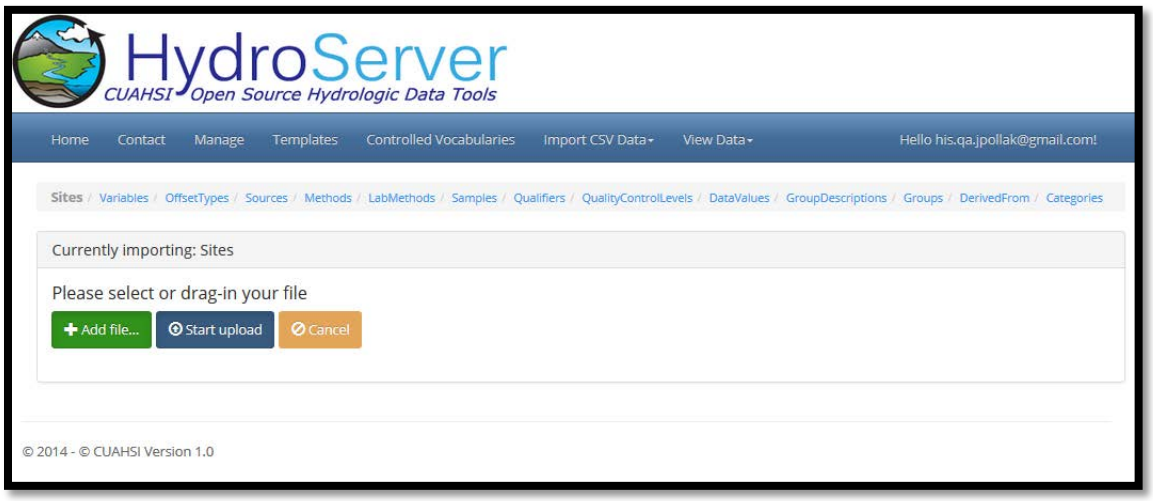
In the left column, there are templates that represent all of the tables that comprise the ODM. Not all of these tables are mandatory, however. For a detailed look at what tables and fields are required, there is an ODM Guide in .xslx format also located on the Templates page on the ODM Uploader for Azure website, which is located on the right side of the page. This file contains a color-coded explanation of mandatory and optional tables, mandatory and optional fields, as well as controlled vocabulary fields. This document was derived from the technical specifications for the ODM.

UPLOADING DATA

To begin uploading templates either click on Import CSV Data into HydroServer on the landing page or click the dropdown for Import CSV Data.



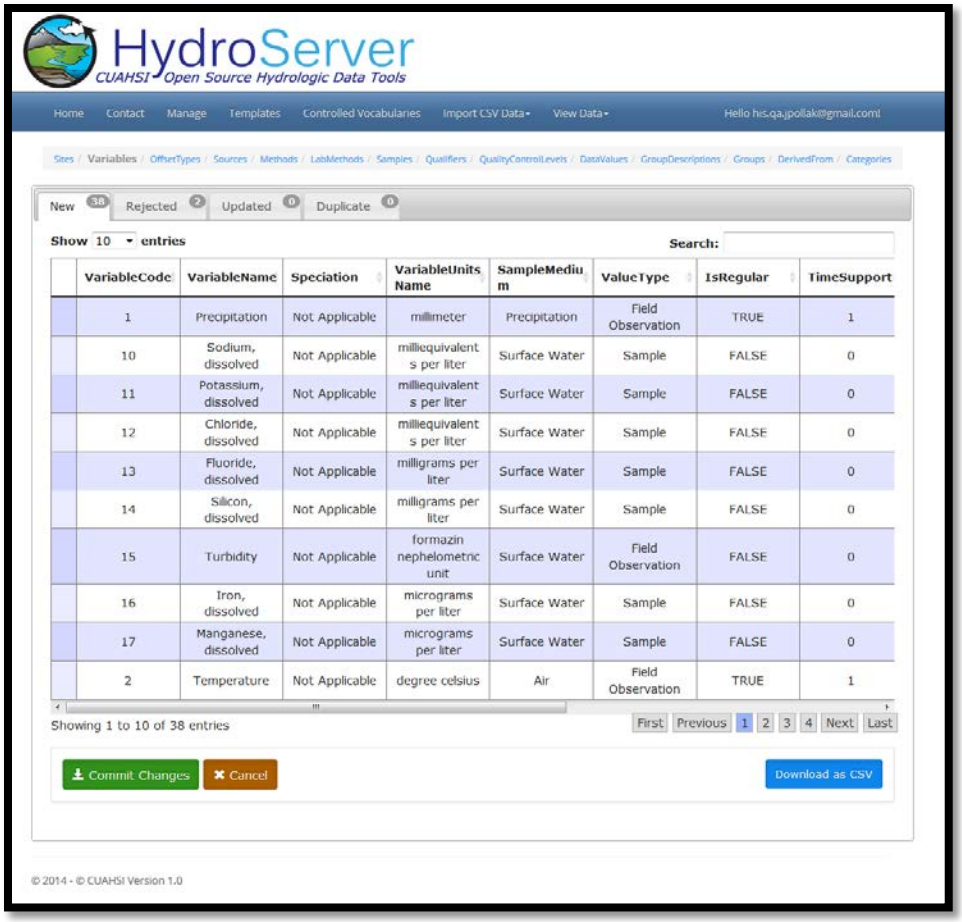
The workflow for uploading templates is an ordered process. Although not all of the templates are required, the upload page for each is listed. From left to right: **Sites**, **Variables**, **OffsetTypes**, **Sources**, **Methods**, **LabMethods**, **Samples**, **Qualifiers**, **QualityControlLevels**, **DataValues**, **GroupDescriptions**, **Groups**, **DerivedFrom**, and **Categories**.



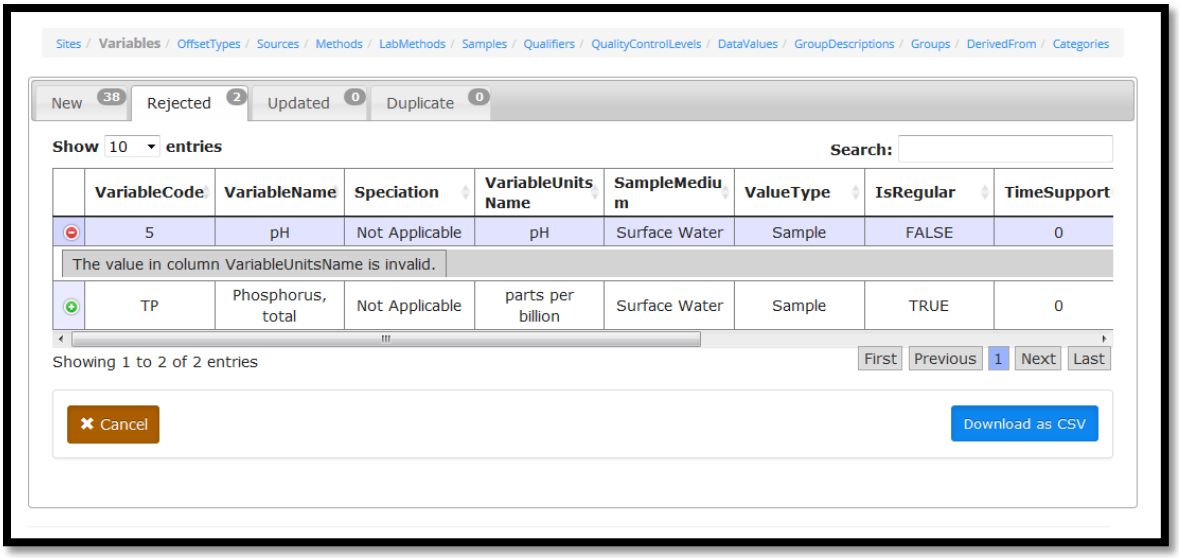
Templates that have been populated with data can be uploaded by going to the corresponding upload page for the template. Begin an upload by either dragging a file on to the *Add File…* button or click *Add File…* and navigate to the location of the .csv file. Clicking *Start Upload* will upload the data to CUAHSI's server for validation.

The application will return a webpage that appears similar to the screenshot below where the rows from your .csv file have been parsed into four categories:

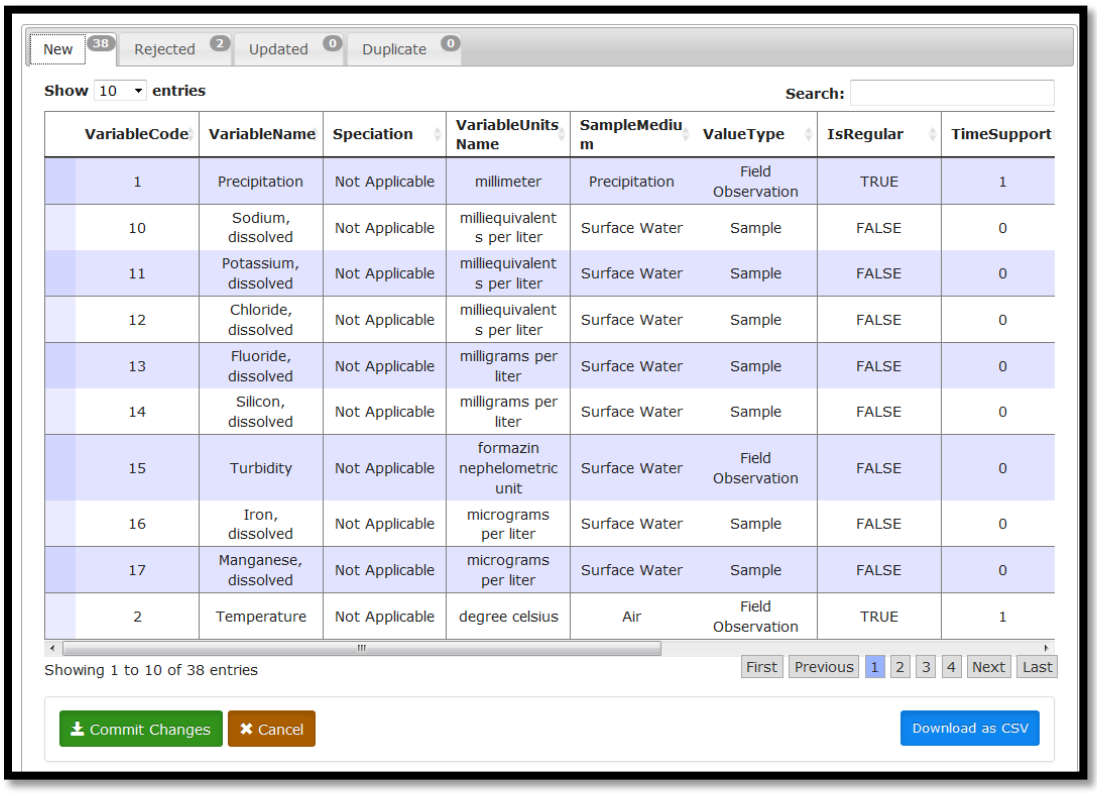
- **New**: Valid records ready to be committed to the database.
- **Rejected**: Records that could not be validated by the application.
- **Updated**: Records that have been updated because the uploaded record had a previously used key with other, changed metadata that are ready to be committed to the database.
- **Duplicate**: Records that are a complete match of records that have been previously committed to the database.



Navigate between the four categories by clicking the labeled tabs. In the *Rejected* tab, view error messages by clicking the green plus symbol in the far left column. Download the rejected records as a .csv file by clicking the *Download as CSV* button on the right as seen below:

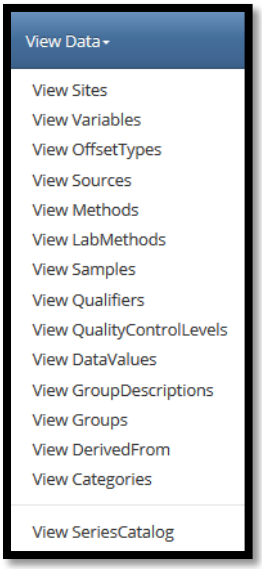


To commit the new and updated records to the database click the *Commit Changes* button, which is located in the lower left of the *New* tab.

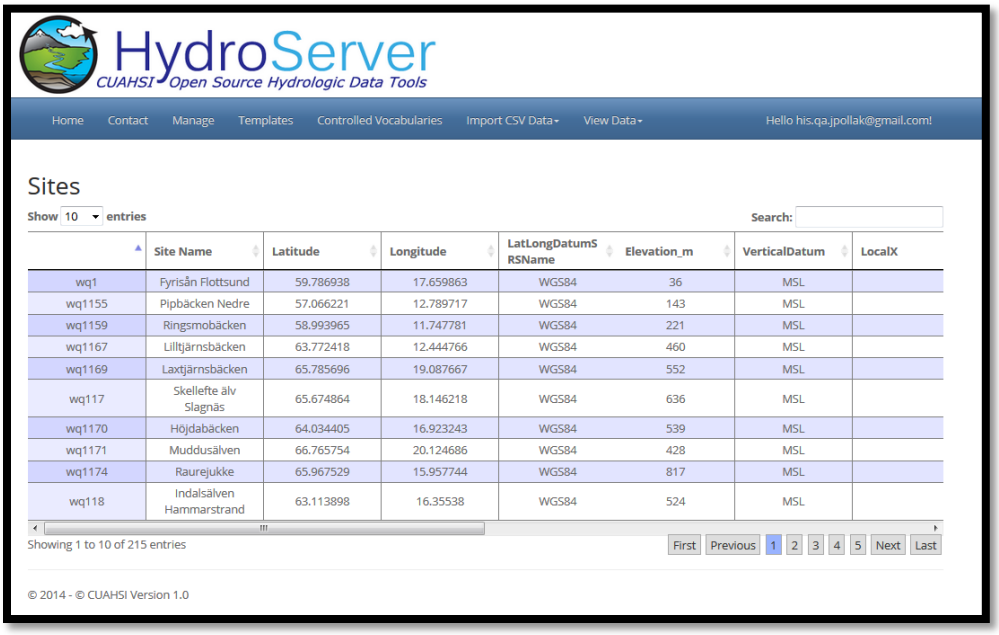


Viewing and Editing Data in the Database

To view data that have already been loaded into the database, click the View Data and select the table that you wish to view.



The resulting webpage will allow you to view, search, and sort through data that have been committed to the database.



Editing the contents of a database requires a new upload. For Sites and Variables, uploading a record with a previously used Code (SiteCode or VariableCode) will update the corresponding record in the database with the information from the new upload. Such records will be presented in the validation step during upload under the *Updated* tab. Other edits must be performed by first deleting an entire table, in the Manage page, then uploading new records.

Registering a Data Source with HIS Central

Registering your data service as a data source in the WDC's HIS Central Catalog requires just two items:

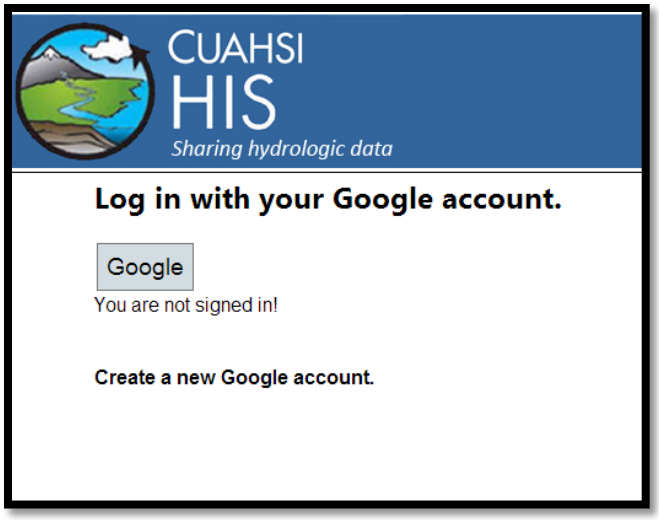
1. A Google account for authentication
2. The URL for the data service (provided by CUAHSI staff)

The following tasks are must be completed in order for data to be discoverable through data access software like HydroDesktop:

1. Authenticate on the HIS Central website (hiscentral.cuahsi.org)
2. Add URL for the data service and applicable metadata
3. Await or request a harvest of the metadata

Authenticating on HIS Central

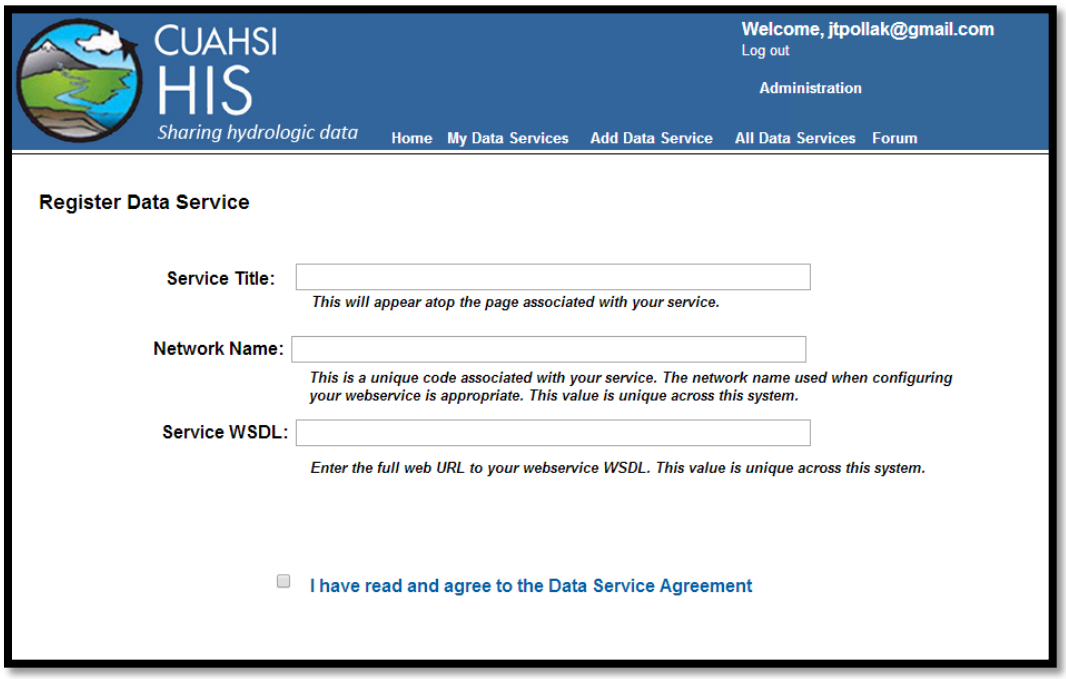
Beginning in 2014, authentication on the HIS Central website was changed to use the OpenID protocol. Consequently, the easiest way to sign on to HIS Central, like the ODM Uploader for Azure website, is by using a Google account.



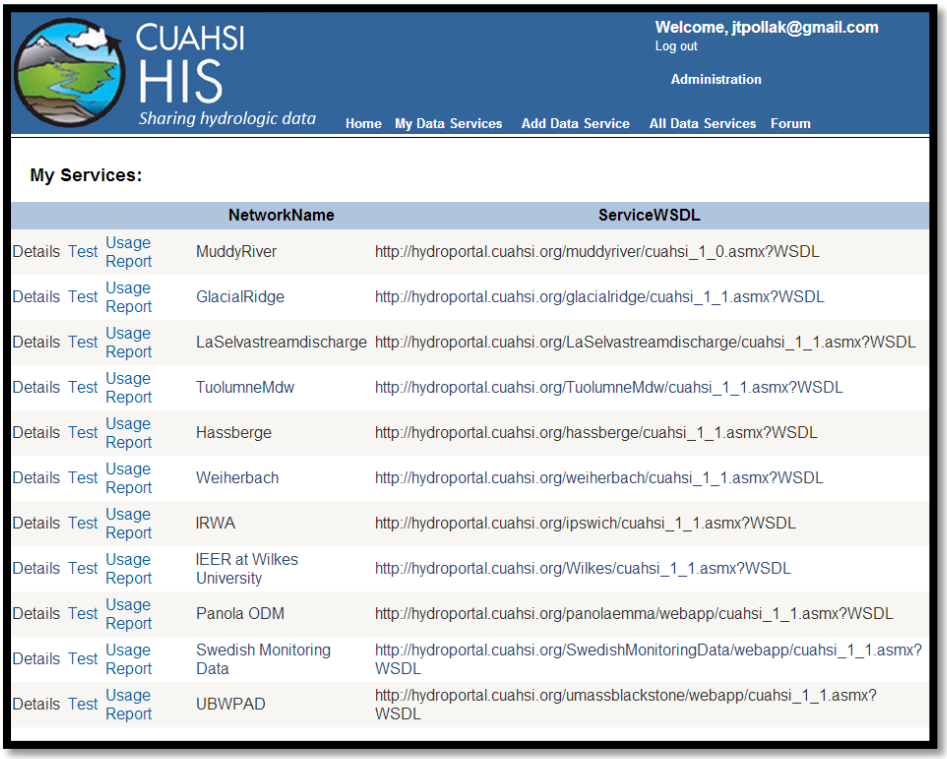
To sign on to HIS Central, visit the website (hiscentral.cuahsi.org), click Login, and follow the prompts to allow HIS Central to access your Google account.

Adding a Data Service and Metadata

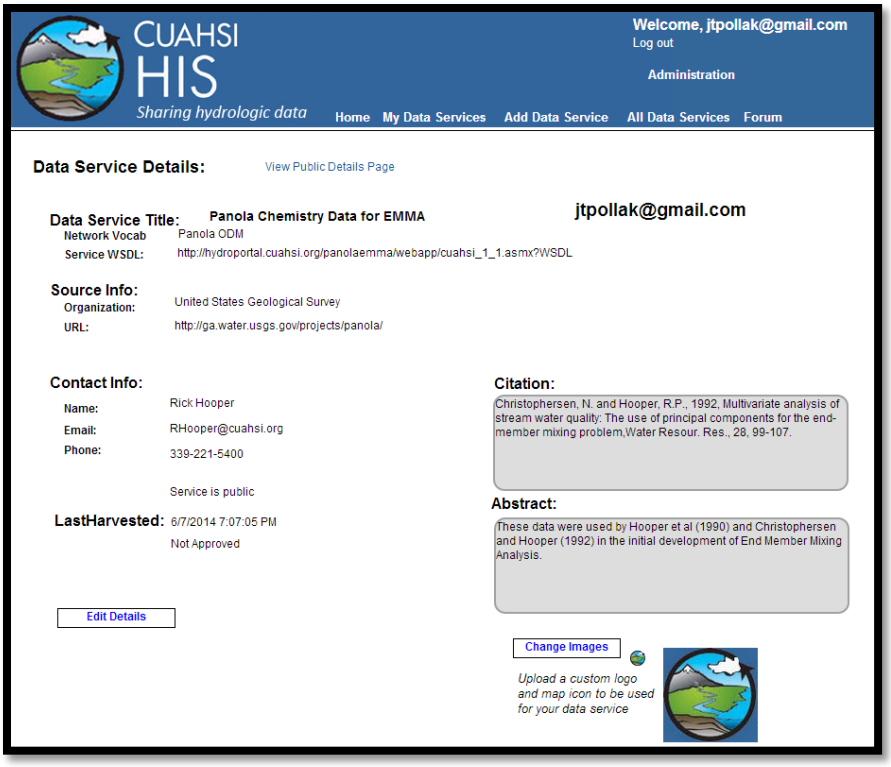
Once signed on to the HIS Central website, it is then possible to add a data service to the catalog. To do this, click Add Data Service, give the data source a Service Title, fill in the Network Name that was used in the database request to CUAHSI, and finally fill in the WSDL URL, which is provided by CUAHSI staff.



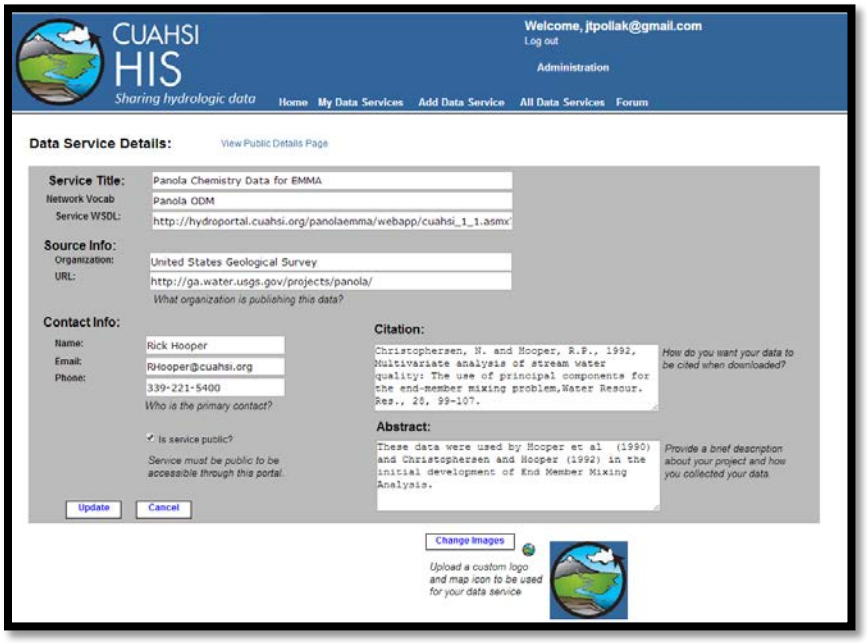
After a new data service is added to the HIS Central catalog, the next step is to add additional metadata about the publishing entity. This can be done by clicking My Data Services tab, then clicking the Details link for a specific service, which can be seen on the left side of the screenshot below.



The following page will show a preview of the metadata for the data publisher. To edit this information, click Edit Details, which is seen in the lower left of the screenshot below.



The subsequent page will allow you to edit information such as contact information for the data publisher:

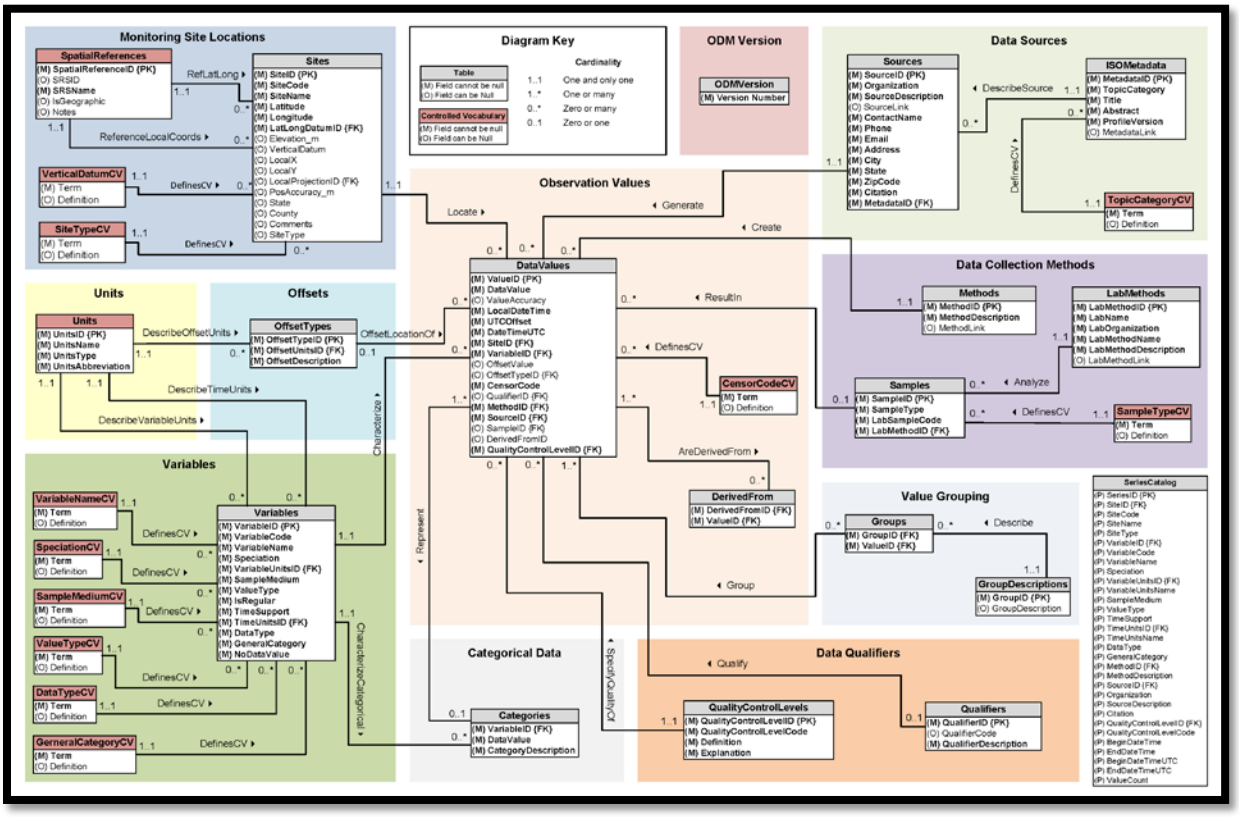


THE HIS CENTRAL METADATA HARVESTER

The HIS Central Catalog utilizes are harvester to gather metadata from each data source registered in the catalog. Much like a web crawler that visits and gathers data from websites, this harvester is a computer program that visits each registered data source and gathers metadata. The metadata collected and stored by the HIS Central enables the search and discovery of data. Data services that are registered are harvested once a week. If a data publisher needs their data source to be harvested prior to the next scheduled harvest they can request it from CUAHSI staff.

Once a data source has been harvested by HIS Central, it will be discoverable by clients that search the catalog like HydroDesktop.

APPENDIX A: THE OBSERVATIONS DATA MODEL (ODM)



The Observations Data Model (ODM) is the information model employed by the CUAHSI HIS. The most common implementation of it is as relational database in Microsoft SQL. The core of the model comprises of a center table that contains the value of observations as well as foreign keys to ancillary tables. These tables provide metadata with the goal of unambiguous interpretation of the data values and include tables with information related to the:

- Location of the observation (*Sites*)
- Phenomenon being observed (*Variables*)
- Methods being employed (*Methods*)
- Sources of the data (*Sources*)
- Quality control techniques employed (*Data Qualifiers*)

The ODM is described in a peer-reviewed article in Water Resources Research and can be accessed here.

For additional information, visit the WDC website.

Appendix B: Required and Optional Templates

Template Name	Template Description
Variables	The Variables template lists the full descriptive information about what variables have been measured. This template is mandatory.
Methods	The Methods template lists the methods used to collect the data and any additional information about the method. This template is mandatory.
Sites	The Sites template provides information giving the spatial location at which data values have been collected. This template is mandatory.
Sources	The Sources template lists the original sources of the data, providing information sufficient to retrieve and reconstruct the data value from the original data files if necessary. This template is mandatory.
Samples	The Samples template gives information about physical samples analyzed in a laboratory. This template is optional.
LabMethods	The LabMethods template contains descriptions of the laboratory methods used to analyze physical samples for specific constituents. This template is optional.
Quality Control	The QualityControlLevels template contains the quality control levels that are used for versioning data within the database. This template is mandatory.
DataValues	The DataValues template contains the actual data values and keys to metadata templates. This template is mandatory.
Categories	The Categories template defines the categories for categorical variables. This template is mandatory when variables exist that have DataType specified as "Categorical." Multiple entries for each VariableCode, with different DataValues provide the mapping from DataValue to category description.
DerivedFrom	The DerivedFrom template contains the linkage between derived data values and the data values that they were derived from. This template is optional.
GroupDescriptions	The GroupDescriptions template lists the descriptions for each of the groups of data values that have been formed. This template is optional and only required if the Groups template is used.
Groups	The Groups template lists the groups of data values that have been created and the data values that are within each group. This template is optional.
Qualifiers	The Qualifiers template contains data qualifying comments that accompany the data. This template is optional.

Appendix C: Controlled Vocabularies in the ODM

Within the ODM, certain fields have been designed as controlled vocabularies, which are community driven and moderated by CUAHSI staff. Click here to visit the landing page for the Master Controlled Vocabulary Registry for ODM 1.1.

The list of fields that are controlled vocabularies can be seen below listed by template and hyperlinked to the appropriate web page to view, submit edits, or submit new terms:

Variables

- VariableName
- Speciation
- VariableUnitsName
- SampleMedium
- ValueType
- TimeUnitsName
- DataType
- GeneralCategory

Sites

- LatLongSRSName
- LocalProjectionSRSName

Sources

- TopicCategory

Samples

- SampleType

DataValues

- CensorCode