CUAHSI

UNIVERSITIES ALLIED FOR WATER RESEARCH

DATA FORMATTING GUIDE:

Data Publication
with the
CUAHSI Water Data Services

Version 2.0

Prepared by Liza Brazil
Adapted from Rick Hooper

March 2018

CONTENTS

DISCLAIMER

The CUAHSI Water Data Services is funded by the National Science Foundation to *provide services to discover, retrieve, and publish water data. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 13-38606. 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 NSF.*

Introduction and Purpose of This Document

The CUAHSI Hydrologic Information System (HIS) enables data uploading (HydroServer) and data discovery (HydroClient). This document describes how to format your data prior to uploading it to your HydroServer. You must format your data to fit with one of the two CUAHSI formatting options: The Standard Format Option and The Advanced Format Option. The Standard Format Option is a subset of the Advanced Format Option. After formatting your data, you should refer to the *CUAHSI Uploading Guide* to proceed with creating a publishing account, uploading data to your HydroServer, and requesting publication to make your data accessible on HydroClient.

Background: Time Series and the Observations Data Model

The CUAHSI Hydrologic Information System (HIS) is designed for sharing *time series data* - observations collected over time at a fixed point, such as a record of stream stage or a series of water-quality observations at a station.[1] The data stored in the CUAHSI HIS follow a specific information model called the *Observations Data Model* (ODM)[2], which defines a metadata profile for water data. This document describes how to fill in the CUAHSI Formatting Template to comply with the ODM metadata profile. CUAHSI offers the Standard and Advanced Formatting Templates. The Standard Formatting Template is organized into six required tables, while the Advanced Formatting Template includes the six required tables and seven optional tables, as described in Appendix B: Optional Tables and Advanced Features of ODM. The optional tables are used for different purposes, such as tracking derived values (e.g., tracking a calculation from the stage recorded by a sensor to the discharge resulting from a rating curve) or for grouping variables that have some logical organization (e.g., a string of thermistors mounted at a buoy in a lake.) Note that if you fill out any of the optional tables, you must proceed with the Advanced Upload Option as the Standard Upload Option will not handle optional tables.

Within the CUAHSI Formatting Templates, a time series is defined as a collection of observations having the same value for the following five metadata attributes:

1. Site (e.g., USGS Site 08158000, Colorado River at Austin, TX)
2. Property measured (e.g., alkalinity)
3. Method (e.g., Gran Titration)
4. Source (the collecting agency, e.g., University of Texas)
5. Quality Control Level (e.g., level of review of the data, e.g., “raw”, or “edited”)

If any of the values of these five attributes differ, then a new time series is created. For example, the raw (unedited) data is a different time series from the data which have been reviewed and approved because they have a different Quality Control Level.

The *CUAHSI Formatting Template* does not have a table for each time series, but rather has a single *data values* table where each observation contains a link to metadata attributes like site, method, etc. A

[1] ODM can be used to store other kinds of data, but this guide focuses only on time series data.

generalization of the CUAHSI Formatting Template which follows the ODM can be seen below in Figure 3.

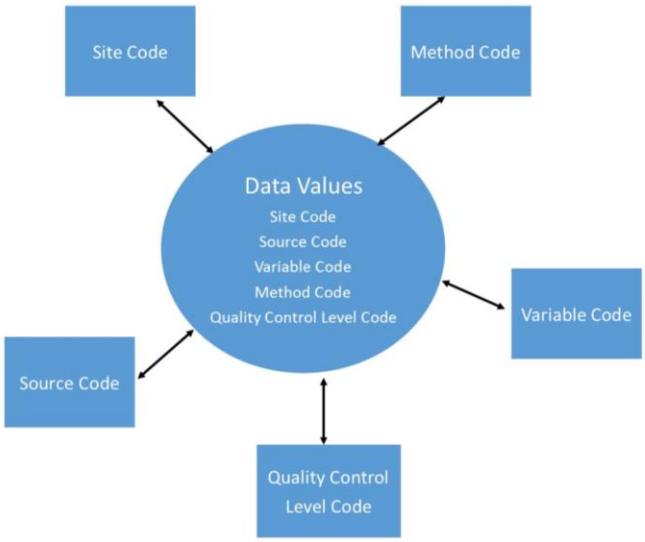


Figure 3. A generalization of the Observations Data Model.

It is important that you fill out all six required metadata tables first and then the data values table. By defining your metadata records before your data values, you will have valid metadata descriptors (referred to as "codes" in the CUAHSI Formatting Templates) for each observation. As you work through this process, it is important to keep a few principles in mind:

- Metadata need only be recorded and uploaded once. Adding additional data values to an existing time series is very easy.
- The codes you use in the metadata tables (site codes, variable codes, method codes, etc.) will be referenced in the data values table.
- For all subsequent data value uploads, you will use the same metadata codes unless a new time series is being uploaded.
- A more complete metadata profile makes your data more useful not only to other scientists, but to your own project members.
- Completing this data publication process fulfills one of your grant requirements, the Data Management Plan, if you received funds from the National Science Foundation and may also fulfill grant requirements from other funders.

See Appendix A for more information on the ODM, including the data schema.

The Example Data Set

As a test case, consider data collected at Panola Mountain Research Watershed located near Atlanta, GA by the USGS that were used as part of a paper describing End Member Mixing Analysis (EMMA), a

multivariate analysis technique[3]. The data consist of measurements of 6 solutes in a set of 905 samples collected at one site over a period of 3 years. The objective of this exercise is to publish this data set for the use of students in a short course on this multivariate technique. Students can discover and download the data using HydroClient or can use the WaterML R Package to perform statistical analysis in any R environment.[4]

The data are in a spreadsheet which contains a Sample ID number (used to track the samples through laboratory analysis), a collection date and time, the values of the stream discharge at the site, and concentrations of six solutes. An excerpt of the data set is shown in Table 1.

Table 1. Example data set from Panola Mountain Research Watershed.

Lab ID Panola	Collect Date	Collect Time	Discharge	Alkalinity	Na ueq	Mg ueq	Ca ueq	H4SiO4 um	SO4 ueq
-94	10/1/1985	10:11	16.16	329.00	153.006	84.725	118.216	325.977	13.699
-60	10/1/1985	10:11	16.16	329.00	152.206	83.1458	120.710	328.972	12.899
-49	10/1/1985	10:54	14.44	328.00	160.406	82.257	116.719	286.341	12.299
-50	10/1/1985	10:59	9.72	330.00	160.006	85.614	114.724	336.960	11.499
-51	10/1/1985	11:11	16.16	322.00	152.606	85.614	116.220	331.668	12.899

These data are typical of many chemical observations in that they are organized by samples where multiple solutes are measured in each sample. Units vary across solutes (some are in microeq/L, others micromol/L) or are not specified (discharge is in L/s, but only project team members would know that). Time is in local time (in this case GMT-5) and standard time is used throughout the year to record time at the site, even though Georgia observes Daylight Savings Time.

Preparing Metadata for the Required Tables

The CUAHSI Formatting Template contains a number of worksheets. The *Introduction* is a brief set of instructions on data uploading. The *Description of Tables* contains a definition of the required and optional tables contained within the ODM. The succeeding tabs are organized by metadata type and contain a complete listing of the fields in the corresponding table, its format, whether the terms used in the field are part of a *controlled vocabulary,* and a link to the associated controlled vocabulary database. A controlled vocabulary is simply a standard list of terms recognized by HIS and used in HydroClient.

Step 1. Defining Properties Measured (Variables)

We start with "what" was measured. In the Panola EMMA data set, we have seven different variables: discharge and the six solutes. Thus this table must ultimately contain 7 rows, one for each variable. The fields in this table are as follows:

[3] Hooper, RP et al. 1989. Modeling streamwater chemistry as a mixture of soil-water end members—an application to the Panola Mountain watershed, Georgia, USA. *J. Hydrol.* **116**: 321-343.

[4] This is an actual data service (http://hiscentral.cuahsi.org/pub_network.aspx?n=385) called "Panola EMMA" and can be discovered and downloaded through HydroClient at http://data.cuahsi.org

1. **Variable Code.** This is a user-defined code that uniquely defines the variable; it is the key that will be used in the data values table to link it with this variable description. If your project already has a unique code defined for each variable, you may use that.[5] If not, you can enter something as simple as a unique number for each. We will number the variables 1 through 7.
2. **VariableName.** *Controlled Vocabulary.* Click on the Controlled Vocabulary hyperlink in the VariableName column. The VariableName CV is a listing of thousands of terms.[6] Use the "search" function of your web browser to locate your term. For example, if we search for the term "alkalinity," the 11 matches of this term arise from five terms (the word "alkalinity" appears in both the term and its definition) and "alkalinity" appears in the definition of the term "pH". Because Gran titration was used at Panola, we choose the term "Alkalinity, total" and enter that into the spreadsheet. The remaining solutes were all measured on filtered samples, so that each solute is modified by the "Dissolved" descriptor as shown below (example: Calcium, dissolved). The term 'Discharge' exists and is used to describe the stream discharge. Note that "streamflow" is another term that exists and is a synonym to discharge. Either may be entered into the spreadsheet. The resulting column of samples names (each listed with a variable code we have defined) is as follows:
    1. Alkalinity, total
    2. Calcium, dissolved
    3. Magnesium, dissolved
    4. Sodium, dissolved
    5. Silica, dissolved
    6. Sulfate, dissolved
    7. Discharge
3. **Speciation.** *Controlled Vocabulary. Default = "Not Applicable".* When some solutes are reported as mg/L, there is a need to specify the speciation of the solute. For example, sulfate ion may be reported as mg/L S (sulfur) or mg/L SO_4 (sulfate). In this case, we are reporting all solutes as micromoles/L or microequivalents/L. Therefore, the CV term "Not Applicable" is used for all rows.

Best Practice: When publishing concentration data, avoid reporting data in mg/L because these data are easily misinterpreted if data users do not read the Speciation field. Use unambiguous units such as micromole/L.

4. **Units.** *Controlled Vocabulary*. Simply choose the units that describe your data. For these data, the following units were chosen (shown with their variable code):
    1. microequivalents per liter
    2. microequivalents per liter
    3. microequivalents per liter
    4. microequivalents per liter
    5. micromoles per liter
    6. microequivalents per liter
    7. liters per second
5. **SampleMedium.** *Controlled Vocabulary. Default value = "Unknown."* This field describes the medium which is being sampled for laboratory analysis or the medium in which a sensor is placed. Because these are samples from a stream, the sample medium is set to "Surface Water" for all variables.

[5] Some restrictions apply to the legal characters that can be used in this field. See the CUAHSI Formatting Templates for details.

[6] If the variable you are measuring does not appear in the CV, you can request that it be added by clicking *New* at the top of the page and submitting a request.

6. **ValueType.** *Controlled Vocabulary. Default value = "Unknown."* This short CV distinguishes between various types of observations, such as model results versus various kinds of observations. In this case, all chemical analyses are the result of a laboratory analysis of a sample; the discharge is a value derived from stage measurements. The following terms were chosen:
    1. Sample
    2. Sample
    3. Sample
    4. Sample
    5. Sample
    6. Sample
    7. Field Observation
7. **IsRegular.** *Boolean (TRUE/FALSE).* This field indicates whether measurements were taken at a regular time step (e.g., every 15 minutes). Simply enter TRUE or FALSE for this field. All of these samples were irregularly spaced so that this field is set to FALSE for all variables.

Note: The discharge samples recorded here are a subset from a continuous time series of discharge (recorded every 5 minutes from a sensor at Panola), but that time series is not published. The discharge used in this exercise is included at the time of sample collection to better describe the conditions under which the samples were collected and is not considered continuous. If the entire discharge time series (every 5 minutes) were to be included instead, then the following would change: IsRegular: TRUE, TimeSupport: 0, and DataType: Continuous.

8. **TimeSupport.** *Default Value = 0.* This numeric field indicates the period of time which the measurement represents. For example, cumulative weekly precipitation would have a time support of 7 days; average daily discharge has a time support of 1 day. In this case, all observations are instantaneous so the time support is set to 0.
9. **TimeUnitsName.** *Controlled Vocabulary. Default Value = "hours".* This field describes the units used in the TimeSupport field (i.e., if weekly discharge TimeSupport was set to "7", units should be "days"). For instantaneous data where time support is set to 0, value is arbitrary. It was set to "minute" in this case.
10. **DataType.** *Controlled Vocabulary. Default Value = "Unknown".* This field describes the recorded value over the time interval being sampled. For instantaneous data, a distinction is made between "continuous" and "sporadic" measurements.
    - "Continuous" data are measured at a sufficiently high frequency that the measurements can be interpreted as a valid representation of the measured phenomenon. A quick theoretical test, is to ask yourself: can a continuous record formed by interpolating between observations be considered valid? If yes, you may label the data type Continuous.
    - "Sporadic" data are collected at a lower frequency (that is, a simple interpolation of observations is *not* a valid representation of the time series).
    - For irregularly spaced data, use the lowest observation frequency to assess the data type. Although subjective, you should record your intent in making the measurement. Generally, sampling frequencies for sensors are set to enable a continuous measurement, but most sensors use some sort of averaging or aggregation over a time support window. For all measurements with TimeSupport values greater than 0, some statistic of the data is recorded over the time support interval. The recorded statistic should be specified as the DataType. For series such as average daily discharge, a higher frequency recording (generally every 15 minutes for large rivers and as frequently as every 1 minute for small streams) is averaged to produce this record. For

some measurements using analog devices, such as precipitation measured weekly with a standard rain gage, the device is "continuous" in the sense that it is open to the atmosphere at all times but only the cumulative amount of rain is recorded. For further information, examine the controlled vocabulary listing. All the data in the Panola data set are considered to be "sporadic" because their lowest frequency (weekly) is much lower than the rate at which the values actually change.

Note: Time support is not the same as spacing. For example, it is possible to have a time series in ODM where values are recorded every 30 minutes (spacing), but where the time support is equal to the period over which the measurement was actually made (e.g., a turbidity sensor that wakes up every 30 minutes and takes 100 instantaneous observations over a 5 second time support window and records the median value).

ODM allows the publication of multiple time series based on the same continuous data series, each with different averaging periods, and enables these series to be linked using the "Derived From" and "Groups" tables. For example, time series of stream stage recorded every minute can be linked to stream discharge derived from stage, as well as time series of daily average and monthly average discharge. See Appendix B, "Advanced Features."

11. **General Category.** *Controlled Vocabulary. Default Value = "Unknown."* This field labels the category of the data such as "Water Quality" or "Hydrology." This provides a convenient grouping of terms for data users. Simply choose the category that best describes your data or set to unknown. For these data, the following terms were used
    1. Water Quality
    2. Water Quality
    3. Water Quality
    4. Water Quality
    5. Water Quality
    6. Water Quality
    7. Hydrology
12. **No Data Value.** Enter the numeric value used to indicate "no data" for your project. At Panola, "-99" is used as this value. In this data set, there are no missing values. This numeric value can be used in the Data Values sheet to represent when there is "no data."

Note: You should pick a no data value that is well outside the plausible range of values for your measurements. A typical no data value is -9999.

STEP 2. DEFINING THE METHODS USED

Now we move to "how" it was measured. This is a much simpler table that consists of only 3 fields.

1. **MethodCode.** Just like Variable Code, this is a unique, user-defined field for each method. If your project has defined method codes, use them; otherwise you can use a unique number to refer to each method. This example uses 1-4 as method codes.
2. **MethodDescription.** A text field to allow you to describe the method used to make the measurement. You may be as descriptive as you want in this field, and you are encouraged to provide enough detail that

a data user can understand how you made the measurement. For the Panola data the following method codes and method descriptions were defined:

1. V-notch weir; stage measured with high res potentiometer.
2. Directly Coupled Plasma (DCP) spectrophotometer
3. Automatic titration system
4. Ion Chromatography (IC)

3. **MethodLink.** An optional text field that allows you to put in a URL to a web page that describes a method more completely. For Panola data, this field was left blank.

STEP 3. DEFINING SITE METADATA

Next we record "where" the measurements were collected. Although this table contains a number of fields, only 4 are mandatory (with additional fields populated with "Unknown"): a Site Code, a Site Name, latitude and longitude. All of the data for the example data set were measured at one site, so there will be only 1 row in the sites spreadsheet.

1. **SiteCode.** A unique, user-defined, text code for the site. We used the number "100" to refer to the site at Panola.
2. **SiteName.** A user-defined name for the site. For the example data set, the site name is "Lower Gage."
3. **Latitude.** Expressed in decimal degrees, this is a number between -90 (the South Pole) and 90 (the North Pole). The latitude of the Lower Gage is 33.6318.
4. **Longitude.** Expressed in decimal degrees, this is a number between -180 and +180 with 0 being the Greenwich meridian and +/-180 being the International Date Line. West Longitude is negative; East Longitude is positive. The longitude of the Lower Gage is -84.1724.
5. **LatLongDatumSRSName.** *Controlled Vocabulary. Default value: "Unknown."* Choose the datum for the earth spatial reference system from this controlled vocabulary or set to "Unknown." This information can be found under "settings" on the GPS used to capture the latitude and longitude of the site. For Panola, the coordinates were captured in the NAD83 datum.
6. **Elevation_m.** *Default value = Null.* Enter the elevation of the site in meters. Left blank for Panola.
7. **VerticalDatum.** *Controlled Vocabulary. Default Value: "Unknown."* Enter the datum for defining the Elevation_m. Left blank for Panola.
8. **LocalX.** *Default value = Null.* The X-coordinate of a local coordinate system, either a formal system, such as State Plane projection or a user-defined coordinate system. Left blank for Panola.
9. **LocalY.** *Default value = Null.* The Y-coordinate of a local coordinate system, either a formal system, such as a State Plane projection or a user-defined coordinate system. Left blank for Panola.
10. **LocalProjectionSRSName.** *Controlled Vocabulary. Default value: "Unknown."* Choose the datum for the earth spatial reference system from this controlled vocabulary or set to "Unknown." Left blank for Panola.
11. **PosAccuracy_m.** *Default value = Null.* Enter the accuracy (expressed as error in meters) of the x,y coordinates in m. Left blank for Panola.
12. **State/Administrative Subdivision 1.** *Default value = Null.* For US, enter the state name; for other countries enter province or other first-level subdivision. Set to "Georgia" for Panola.
13. **County/Administrative Subdivison 2.** *Default value = Null.* For US, enter the county name; for other countries, enter second-level subdivision. Set to "Rockdale" for Panola.
14. **Comments.** *Default value = Null.* Open text field to record any comments about site.
15. **SiteType.** *Controlled Vocabulary. Default value = Null.* This site classification comes from the USGS system and is intended to provide a grouping for similar sites, such as wells or stream sites. Choose the appropriate value from the CV. This has been set to "Stream" for the Panola EMMA dataset.

STEP 4. DEFINING THE COLLECTING ORGANIZATION METADATA ("SOURCE")

Now we describe "who" measured or created the data. In many cases, there will be only one entry in this table, but there are situations where there may be multiple entries. For example, there may be a single publication data service for a field site contributed by multiple entities such as universities and government agencies. Alternately, the data service may contain data aggregated from a network of groups, such as the Shale Network[7] in Pennsylvania, which assembles data collected by multiple citizen organizations into a single data service to provide a larger-scale picture of water quality conditions across a region where fracking is underway.

This table contains fields that complete a standard geographic metadata profile (ISO19115).

> **Note:** When the data service is registered with HISCentral, each record in the source table can be associated with multiple grants. This provides a means to demonstrate to the funding agency that your data have been published. This will be a requirement for some grants from the National Science Foundation.

1. **SourceCode.** Enter a unique text code for each organization. In general, just number organizations from 1 to the number of organizations collecting data. For the Panola Data set, "1" was entered. Alternatively, if an organization has an acronym, use that (e.g., "USGS" for United States Geological Survey).
2. **Organization.** Enter the name of the organization responsible for the data collection. This would typically be a university, government agency, or non-governmental organization (like a watershed association). Generally use simply the organization name (e.g., "Utah State University") rather than any department or subunit. This information is the single most important field for data users to assess the reliability of the data. For the Panola data, this is simply "United States Geological Survey" because USGS protocols were followed for data collection, lab analyses, and data processing, rather than a more narrow designation such as the office of the USGS that collected these data.
3. **SourceDescription.** Describe the project or other subunit of the organization more completely. For Panola, we listed the collaborating organizations, and project number.
4. **SourceLink.** Enter a URL that describes the project. For Panola, we entered the web site of the Panola Project.
5. **ContactName.** Enter the name of the person who should be contacted if the user has questions about the data. Generally, a long-term employee (rather than a student) should be the contact. For the Panola Data set, Rick Hooper was listed because he knows this data set, even if he is no longer with the US Geological Survey.
6. **Phone.** Enter the phone number of the contact. Rick's phone number is entered for the Panola data set.
7. **Email.** Enter the email address of the contact. rhooper@cuahsi.org is entered for Rick.
8. **Address.** Enter the street address of the contact. The CUAHSI office address is listed for the Panola data set.
9. **City.** Enter the city of the contact. "Medford" for the example data.
10. **State.** Enter the state of the contact. "MA" for the example data.
11. **ZipCode.** Enter the zip code/postal code for the contact. "02155" for the example data.
12. **Citation.** Enter how you wish the data to be cited. This can be to a paper in which the data appear or could be another citation that you wish to track. A journal citation was entered for the Panola data set.

[7] For more information about the Shale Network, see: http://www.shalenetwork.org

13. **TopicCategory.** *Controlled Vocabulary. Default = "Unknown."* These categories arise from an international standard—ISO19115—to categorize topics. Choose the one that fits your data set best. For Panola, we used "Inland Waters".
14. **Title.** Provide a title for your data set. "Panola Chemistry Data for EMMA" was entered for the example data.
15. **Abstract.** Provide an abstract that describes the data set in more detail. In particular this is an opportunity to describe why the data were collected and to describe the project more fully. This abstract will appear on a web page describing the data service.
16. **ProfileVersion.** If your organization follows a specific metadata profile (especially if it differs from ISO19115 which ODM is compliant with) enter that profile here. Otherwise, set to "Unknown." This field was set to "Unknown" for Panola data set.
17. **Metadata Link.** Enter the URL of a web page that further describes and provides additional metadata for your data. Set to "Null" for Panola data set.

STEP 5. DEFINING THE QUALITY CONTROL LEVEL

When publishing data for others to use, it is important to let the data user know whether the data have been reviewed—in essence, "how good" are the data? You can use Table 2. if the descriptions fit. If you do use the codes and descriptions (Table 2.), copy and paste them exactly into your ODM sheet.

Table 2. Typical definition of QCLs for sensor data

QCL Code	Definition
0	Raw (unedited) data, such as the output from a sensor
1	Data that have passed some objective quality assurance procedure (edited data)
2	Derived product, potentially from multiple sensors, that require some subjective judgment (e.g., inference of a basin average precipitation from multiple rain gages)
3	Interpreted product, requiring more assumptions and judgments (e.g., rainfall inferred from radar reflectance)
4	Knowledge product, requiring inferences, potentially from multiple disciplines, and combining multiple measurements, each with their own assumptions (e.g., proportion of "new" and "old" water in streamflow inferred from stable isotopes)
-9999	Unknown

This convention is typically used for sensor data. However, you can define any set of QCL codes that reflect how your data were quality assured and enter that information with the QualityControlLevels template.
For Panola, the QCLs listed in Table 2 were used.

Best Practice. When publishing sensor data, publish both the raw data (QCL=0) and the edited data (QCL=1) to enable other scientists to see what the sensor actually recorded. What one scientist considers noise, may be signal to another scientist. Additionally, knowing the level of editing that was performed on the raw data to create the final product may be useful in determining whether a dataset is fit for a particular analysis.

ENTERING THE DATA VALUES

With these metadata values defined, you are ready to enter the data values themselves so long as your data conform to the following conditions:

1. Data are recorded at a single site; no "offset" values apply where data are recorded at some distance relative to a fixed point, such as a string of thermistors.
2. Data values are unqualified; that is, you do not wish to mark individual observations with some code. Generally, qualifiers are used to indicate special conditions that the data user should be aware of and may affect the accuracy or the interpretation of the measurements. Examples are measurements made under ice, incomplete daily totals for a measurement, or holding time exceeded for a laboratory measurement. Note that censored observations, e.g., chemical analyses that were performed yet could not quantify a concentration or sensors that were overloaded, are recorded using the *CensorCode* controlled vocabulary which is a different field located in the *Variables* table.
3. You do not wish to record laboratory information about samples. ODM allows one to record laboratory methods, sample codes, and other lab information. For the Panola data set, we chose not to record this information.
4. Data are continuous and not categorical. If data are categorical, the categories must be defined before data are uploaded.

If you wish to use any of these features described above, consult Appendix B, Optional Tables. As with other metadata fields it is easier to upload your data if you define any of these features prior to uploading the data values.

ODM has additional documentation features, described in Appendix B, that allow users to record sensor data, such as stream stage, and define derived values such as average daily streamflow. Different logical groupings can also be defined as needed. See Appendix B for further information on these advanced features. These features can be added after data values are uploaded.

POPULATING THE DATAVALUES TABLE

It will be easiest to fill out the DataValues table if we consider each of the seven time series contained in Table 1 separately. We will lay out a step-by-step approach for one time series (alkalinity); the other time series will be entered in a similar manner. You may stack each time series in the DataValues sheet or you may separate each time series into a new data values sheet.

1. **Duplicate "DataValues" worksheet.** Right-click (PC) or Apple-click (Mac) the "DataValues" tab and Copy the worksheet. Rename it "Alkalinity."
2. **Transfer "Alkalinity" column.** Copy all the values in the Alkalinity column to the "DataValue" column (Column B).
3. **Transfer the Date/Time.** Combine the Date and Time into a single column by adding the values together. Choose the number format to display both date and time and validate that combined date/time data is accurate. Copy the combined date/time field into the "LocalDateTime" column (Column D). Be careful to retain alignment between DataValues and LocalDateTime. Validate that the data are transferred accurately from the original spreadsheet and are in the format MM/DD/YYYY (American date format).

4. **Enter UTC Offset.** Enter the number of hours between the local time and UTC in Column E. For Panola, this is -5 hours (i.e., 5 hours behind UTC). Times are recorded only in standard time, so this is a constant throughout the data set. Copy this value for all data rows.
5. **Calculate DateTimeUTC.** Enter formula in Column F to add LocalDateTime to UTCOffset to calculate DateTimeUTC. Copy this formula for all data rows.
6. **Enter SiteCode.** All samples in the Panola data set were collected at Site "100". Enter "100" in Column G and copy this value for all data rows. This will ensure that the data are linked to the correct Site when they are loaded.
7. **Enter Variable Code.** Consult the Variable table to determine that the VariableCode for Alkalinity is "1". Enter "1" in Column H and copy for all data rows.
8. **Enter the CensorCode.** *Controlled Vocabulary. Default value = "nc."* If none of the data values are censored in the data set, simply enter "nc" (which stands for "not censored") and copy for all data rows. This was done for the Panola data set. If some data values are censored, then enter the appropriate censor code from the CensorCode CV to the appropriate data row.
9. **Enter the MethodCode.** Choose the code that corresponds to the measurement method used for these data in the Methods table. For the alkalinity data at Panola, the method code is "3". Enter "3" into Column M. Copy for all data rows. Note that if the method is unknown, one can enter "0". However, this is not recommended.
10. **Enter the SourceCode.** Choose the SourceCode that corresponds to the collecting organization. For Panola, enter "1" into Column N for all rows.
11. **Enter the QCLCode.** Choose the QCLCode from the QCL table that corresponds to the status for this data series. For Panola, all the data have a QCL Code of "1". Enter "1" into Column Q and copy for all data rows.

This completes all mandatory fields for the Alkalinity data. There are a few optional fields that merit some attention.

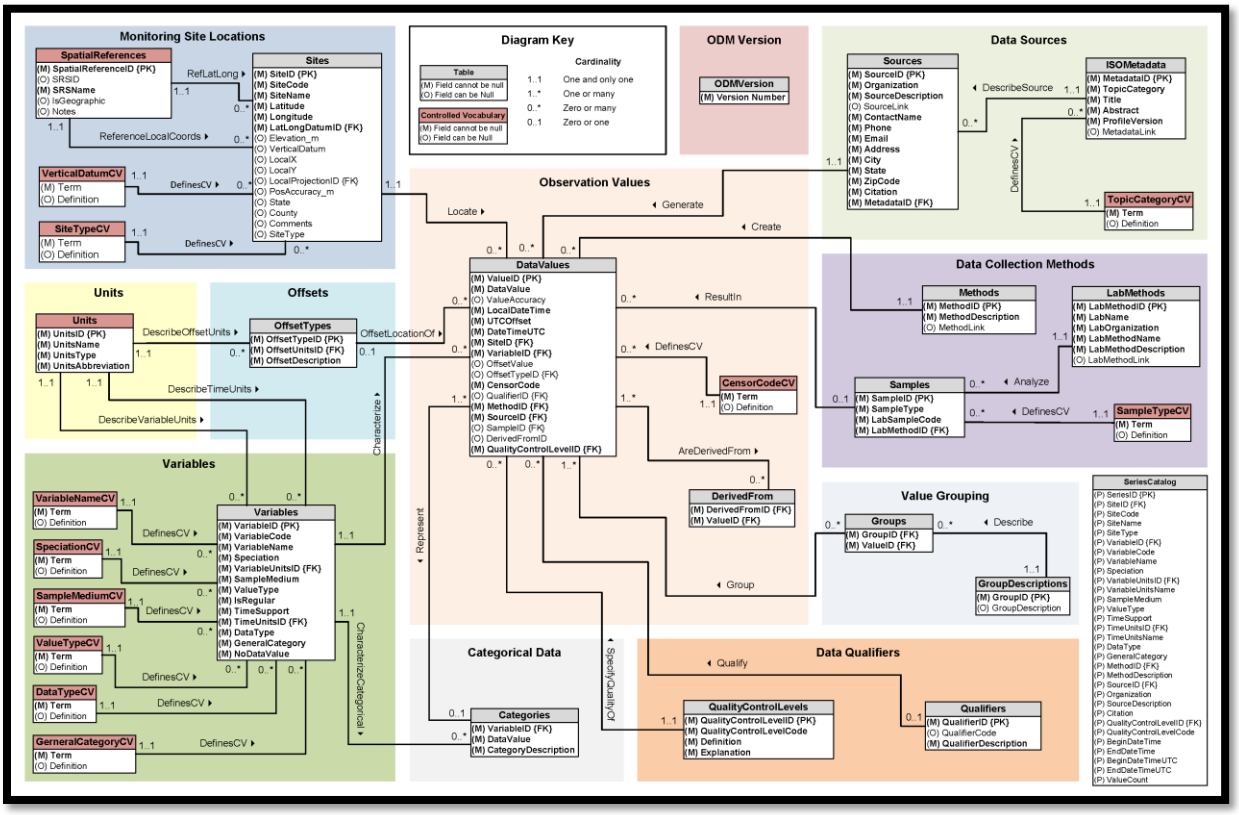
12. **ValueAccuracy.** Formally, this optional field is a numeric value that combines both bias and precision estimates by using a root-sum-square. The suggested use of this field, however, is to report unbiased estimates of the measurement (either through correcting known biases in a new corrected time series or by assuming measurements are unbiased) and reporting simply the precision of the measurements in this field. Providing some guidance to the data users of the data accuracy is very useful and should be included. For the Panola data set, ValueAccuracy was left blank.

All other optional data fields should be left blank. The same series of steps should be repeated for each of the remaining 6 time series.

EXPORTING THE WORKSHEETS

Each tab of the CUAHSI Formatting Template must be exported as a .csv file for upload to your HydroServer. It is recommended that all .csv files containing more than 500,000 rows be zipped to reduce the file size. Before exporting as a .csv, remember to delete rows 2 through 5 and Column A in the templates that contain documentation. The spreadsheet should contain only a single header row followed by all of the data rows. Export each worksheet as a .csv file (use "Save As" and choose the Text (CSV) option). For larger .csv files right click on the .csv files, scroll over "Send to" and select "Compressed (zipped) folder". You are now ready to upload the data. Consult the *CUAHSI Uploading Guide* for uploading instructions.

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 Server. 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 to create the observations (*Methods*)
- Sources of the data (*Sources*)
- Quality control techniques employed (*QualityControlLevels*)

The ODM is described in a peer-reviewed article in Water Resources Research and can be accessed here[8].

[8] DOI: 10.1029/2007WR006392

APPENDIX B: OPTIONAL TABLES AND ADVANCED FEATURES OF ODM

The following tables are optional and not required for the most common data published in the HIS. For more information, see the *ODM Specifications*[9] document or contact help@cuahsi.org.

Samples	The Samples table gives information about physical samples analyzed in a laboratory.
LabMethods	The LabMethods table contains descriptions of the laboratory methods used to analyze physical samples for specific constituents.
Categories	The Categories table defines the categories for categorical variables. This table 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 table contains the linkage between derived data values and the data values that they were derived from.
GroupDescriptions	The GroupDescriptions table lists the descriptions for each of the groups of data values that have been formed.
Groups	The Groups table lists the groups of data values that have been created and the data values that are within each group.
Qualifiers	The Qualifiers table contains data qualifying comments that accompany the data.

[9] DOI: 10.1029/2007WR006392