CUAHSI Community Observations Data Model
Working Design Specifications Document – Version 3[1]

June 23, 2006

David G. Tarboton[2], Jeffery S. Horsburgh[2], David R. Maidment[3], and Blair Jennings[4]

Abstract

The CUAHSI Hydrologic Information System project is developing information technology infrastructure to support hydrologic science. One aspect of this is a data model for the storage and retrieval of hydrologic observations in a relational database. The purpose for such a database is to store hydrologic observations data in a system designed to optimize data retrieval for integrated analysis of information collected by multiple investigators. It is intended to provide a standard format to aid in the effective sharing of information between investigators and to allow analysis of information from disparate sources both within a single study area or hydrologic observatory and across hydrologic observatories and regions. The observations data model is designed to store hydrologic observations and sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used. A relational database format is used to provide querying capability to allow data retrieval supporting diverse analyses. A generic template for the observations database is presented. This is referred to as the Observations Data Model (ODM).

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) is an organization representing more than 100 universities, and is sponsored by the National Science Foundation to provide infrastructure and services to advance the development of hydrologic science and education in the United States. The CUAHSI Hydrologic Information System (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. It is intended to expand the ability for data analysis by providing a standard format to share data among investigators and to facilitate analysis of information from disparate sources both within a single study area or hydrologic observatory and across hydrologic observatories and regions. The hydrologic observations data model is designed to store hydrologic observations with sufficient ancillary information (metadata) about the data values to provide traceable heritage from raw measurements to usable information allowing them to be unambiguously interpreted and used.

---

[1] Version 1 of this design was presented as Chapter 6 in Maidment, D. R., ed. (2005b), 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. Version 2 was circulated by CUAHSI on April 10, 2006.

[2] Utah Water Research Laboratory, Utah State University

[3] Center for Research in Water Resources, University of Texas at Austin

[4] San Diego SuperComputer Center, University of California at San Diego

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.

This design has evolved from an initial design presented at a CUAHSI workshop held in Austin during March, 2005 (Maidment, 2005a) that was then widely reviewed with comments being received from 22 individuals (Tarboton, 2005). These reviews served as the basis for a redesign that was presented at a CUAHSI workshop in Duke during July, 2005 and presented as part of the CUAHSI HIS status report (Horsburgh et al., 2005). Following this presentation of the design, the data model was reviewed and commented on by a number of others, including the CLEANER (Collaborative Large-scale Engineering Analysis Network for Environmental Research) cyberinfrastructure committee. A version 2 Hydrologic Observations Data Model was circulated in April 2006 (Tarboton et al., 2006). This consolidated changes to that point in one document and clarified the generic structure of the model as a conceptual basis for specific software implementations. This document is version 3 of this data model design. The fundamental design has not changed since the status report presentation of the model (Horsburgh et al., 2005) but many table and field names have been changed. Tables have also been added to give spatial reference information, metadata information, and to define controlled vocabularies.

The ODM uses a relational database format to allow for ease in querying and data retrieval in support of a diverse range of analyses. Reliance on databases and tables within databases also provides the capability to have the model scalable from the observations of a single investigator in a single project through the multiple investigator communities associated with a hydrologic observatory and ultimately to the entire set of observations available to the CUAHSI community. The ODM is focused on observations made at a point. A relational database model with individual observations recorded as individual records (an atomic model) is chosen to provide maximum flexibility in data analysis through the ability to query and select individual observation records. This approach carries the burden of record level metadata, so it is not appropriate for all variables that might be observed. For example, individual pixel values in large remotely sensed images or grids are inappropriate for this model.

This data model is presented as a generic template for a point observations database, without reference to the specific implementation in a software system. This is done so that the general design is not limited to any specific proprietary software, although we expect that implementations will take advantage of capabilities of specific software. It should be possible to implement the ODM in a variety of relational database management systems, or even in a set of text tables or variable arrays in a computer program. However, to take full advantage of the relationships between data elements, the querying capability of a relational database system is required. By presenting the design at a general conceptual level, we also avoid implementation specific detail on the format of how information is represented. See the discussion of Dates and Times under ODM features below for an example of the distinction between general concepts and implementation specific details.

Hydrologic Observations

Many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels, and precipitation. National databases such as USGS' National

Water Information System (NWIS) and USEPA's data Storage and Retrieval (STORET) system contain a wealth of data, but, in general, these national data repositories have different data formats, storage, and retrieval systems, and combining data from disparate sources can be difficult. The problem is compounded when multiple investigators are involved (as would be the case at proposed CUAHSI Hydrologic Observatories) because everyone has their own way of storing and manipulating observational data. There is a need within the hydrologic community for an observations database structure that presents observations from many different sources and of many different types in a consistent format.

Hydrologic observations are identified by the following fundamental characteristics:

- The location at which the observations were made (space)
- The date and time at which the observations were made (time)
- The type of variable that was observed, such as streamflow, water surface elevation, water quality concentration, etc. (variable)

These three fundamental characteristics may be represented as a data cube (Figure 1), where a particular observed data value (D) is located as a function of where it was observed (L), its time of observation (T), and what kind of variable it is (V), thus forming D(L,T,V).

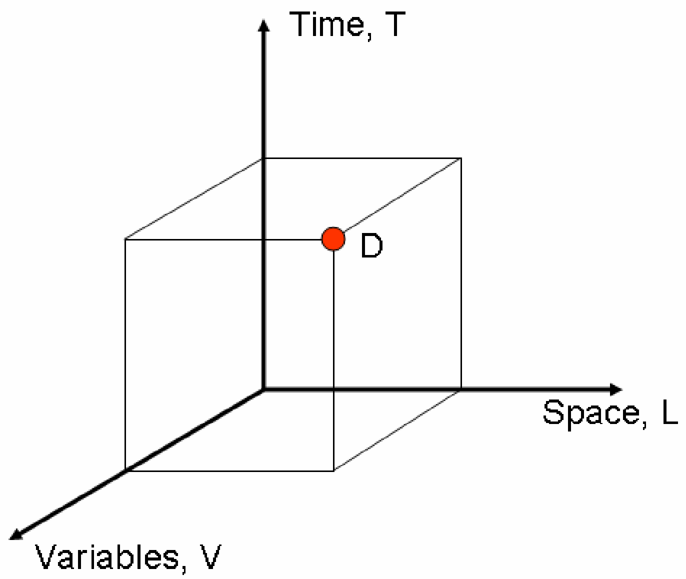


Figure 1. A measured value (D) is indexed by its spatial location (L), its time of measurement (T), and what kind of variable it is (V).

In addition to these fundamental characteristics, there are many other distinguishing attributes that accompany observational data. Many of these secondary attributes provide more information about the three fundamental characteristics mentioned above. For example, the location of an observation can be expressed as a text string (i.e., "Bear River Near Logan"), or as latitude and longitude coordinates that accurately delineate the location of the observation. Other attributes can provide important context in interpreting the observational data. These include data qualifying comments and information about the organization that collected the data. The fundamental design decisions associated with the ODM involve choices as to how much supporting information to include in the database and whether to store (and potentially repeat) this information with each observation or save this information in separate tables with key fields used to logically associate observation records with the associated information in the ancillary tables. Table 1 presents the general attributes associated with a point observation that we judged should be included in the generic ODM design.

Table 1. ODM attributes associated with an observation

Attribute	**Definition**
Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
DateTime	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
Variable	The physical quantity that the value is measuring (e.g. streamflow, precipitation, water quality)
Location	The location of the observation (e.g. latitude and longitude)
Units	The units (e.g. m or m^3/s) and unit type (e.g. length or volume/time) associated with the variable
Interval	The interval over which each observation was collected or implicitly averaged by the measurement method and whether the observations are regularly recorded on that interval
Offset	Distance from a reference point to the location at which the observation was made (e.g. 5 meters below water surface)
OffsetType/ Reference Point	The reference point from which the offset to the measurement location was measured (e.g. water surface, stream bank, snow surface)
Data Type	An indication of the kind of quantity being measured (e.g. an instantaneous or cumulative measurement)
Organization	The organization or entity providing the measurement
Censoring	An indication of whether the observation is censored or not
Data Qualifying Comments	Comments accompanying the data that can affect the way the data is used or interpreted (e.g. holding time exceeded, sample contaminated, provisional data subject to change, etc.)
Analysis Procedure	An indication of what method was used to collect the observation (e.g. dissolved oxygen by field probe or dissolved oxygen by Winkler Titration) including quality control and assurance that it has been subject to
Source	Information on the original source of the observation (e.g. from a specific instrument or investigator 3rd party database)
Sample Medium	The medium in which the sample was collected (e.g. water, air, sediment, etc.)
Value Category	An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation

Observations Data Model

The schema of the Observations Data Model is given in Figures 2 and 3. Appendix A gives details of each table and each field in this generic data model schema. The primary table that stores point observation values is the Values table at the center of the schema in Figure 2. Logical relationships between fields in the data model are shown and serve to establish the connectivity between the observation values and associated ancillary information. Details of the relationships are given in Table 2. Figure 3 gives the controlled vocabulary tables that are not formally linked to other tables, but serve to specify the terms that may be used for the fields where the vocabulary is controlled. A Microsoft Access database file ODMSchema.mdb that implements this schema is provided as an electronic appendix to this document.

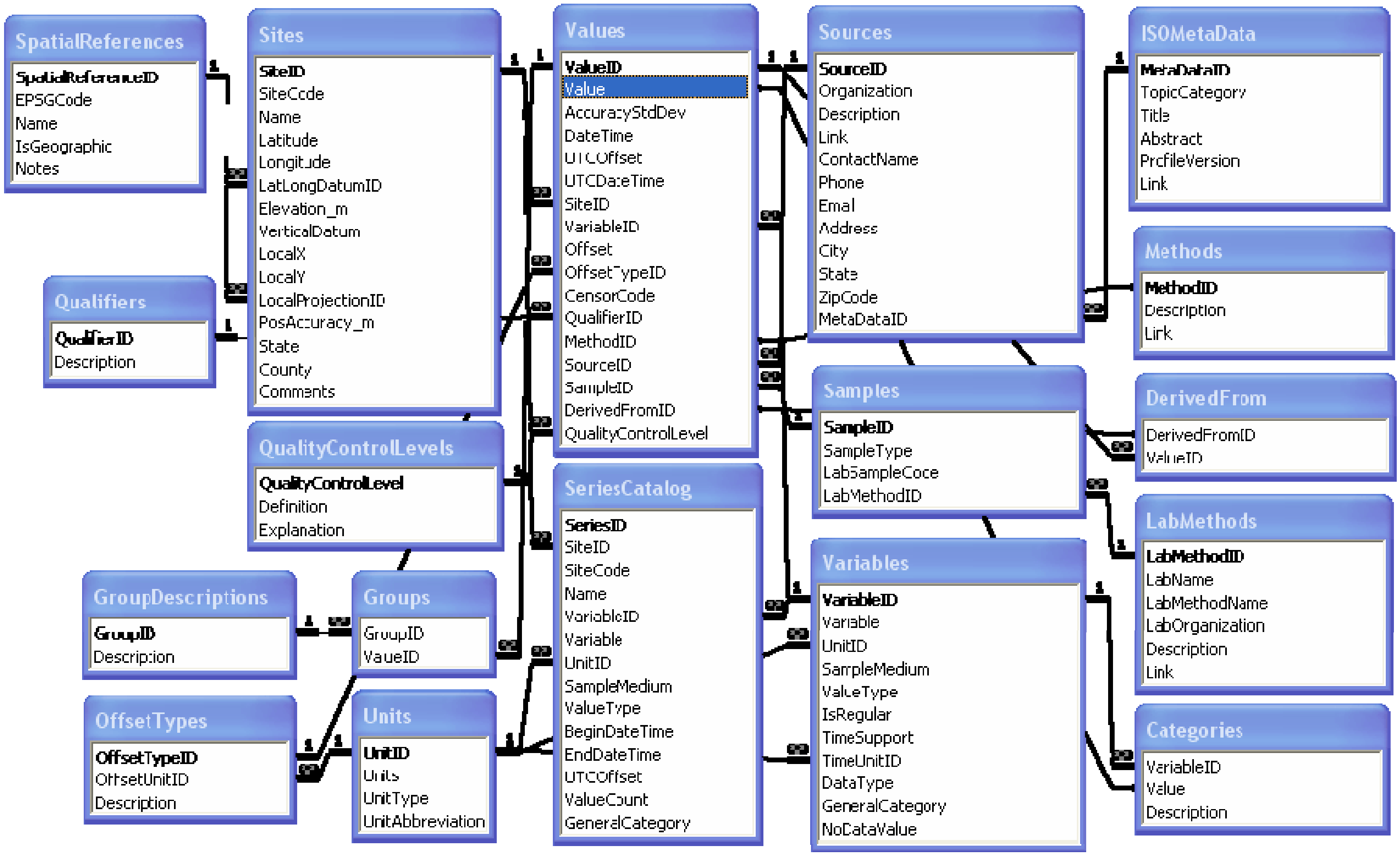


Figure 2. Observations Data Model Schema. Fields in bold are primary key fields.

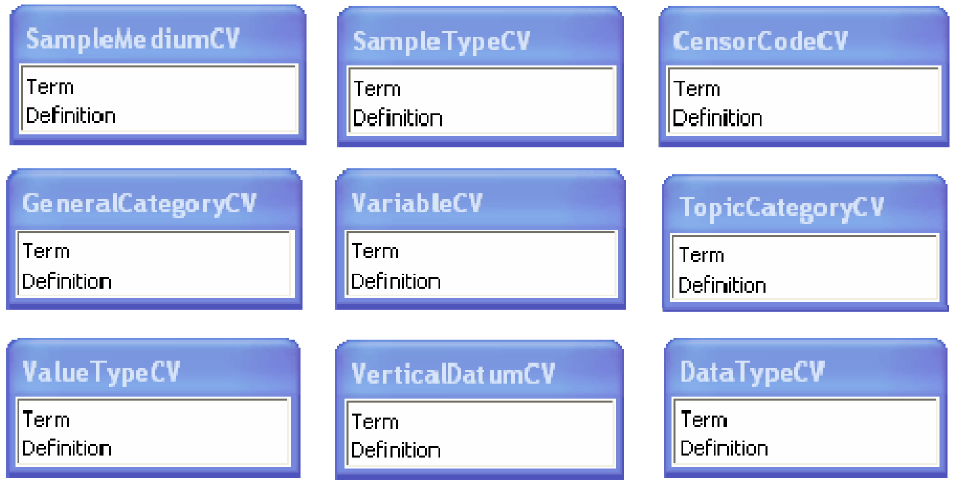


Figure 3. Unlinked Controlled Vocabulary Tables.

Table 2. Observations Data Model Logical Relationships

Relationships that define ancillary information about Values

Table	Field	Type	Field	Table
Values	SiteID	* <-> 1	SiteID	Sites
Values	VariableID	* <-> 1	VariableID	Variables
Values	OffsetTypeID	* <-> 1	OffsetTypeID	OffsetTypes
Values	QualifierID	* <-> 1	QualifierID	Qualifiers
Values	MethodID	* <-> 1	MethodID	Methods
Values	SourceID	* <-> 1	SourceID	Sources
Values	SampleID	* <-> 1	SampleID	Samples
Values	QualityControlLevel	* <-> 1	QualityControlLevel	QualityControlLevels

Relationships that define derived from groups

Table	Field	Type	Field	Table
Values	DerivedFromID	* <-> *	DerivedFromID	DerivedFrom
Values	ValueID	1 <-> *	ValueID	DerivedFrom

Relationships that define groups

Table	Field	Type	Field	Table
Values	ValueID	1 <-> *	ValueID	Groups
GroupDescriptions	GroupID	1 <-> *	GroupID	Groups

Relationships used to define categories for categorical data

Table	Field	Type	Field	Table
Variables	VariableID	1 <-> *	VariableID	Categories
Values	Value	* <-> 1	Value	Categories

Relationships used to define the SeriesCatalog

Table	Field	Type	Field	Table
SeriesCatalog	SiteID	* <-> 1	SiteID	Sites
SeriesCatalog	VariableID	* <-> 1	VariableID	Variables

Relationships used to define the Units

Table	Field	Type	Field	Table
Units	UnitID	1<->*	UnitID	SeriesCatalog
Units	UnitID	1<->*	UnitID	Variables
Units	UnitID	1<->*	TimeUnitID	Variables
Units	UnitID	1<->*	OffsetUnitID	OffsetTypes

Relationship used to define the Sample Laboratory Methods

Table	Field	Type	Field	Table
LabMethods	LabMethodID	1<->*	LabMethodID	Samples

Relationship used to define the Spatial References

Table	Field	Type	Field	Table
SpatialReferences	SpatialReferenceID	1<->*	LatLongDatumID	Sites
SpatialReferences	SpatialReferenceID	1<->*	LocalProjectionID	Sites

Relationship used to define the ISOMetaData

Table	Field	Type	Field	Table
IsoMetaData	MetadataID	1<->*	Sources	MetadataID

Relationship type is indicated as One to One (1<->1), One to Many (1<->*), Many to One (*<->1) and Many to Many (*<->*). The first set of relationships defines the links to tables that contain ancillary information. They are used so that only compact (integer) identifiers are stored with each value and thus repeated many times while the more voluminous ancillary information is stored to the side and not repeated. The second set of relationships defines derived from groupings used to specify values that have been used to derive other values. The third set of relationships defines logical groupings of values. The fourth set of relationships is used to specify the categories associated with categorical variables. The fifth set of relationships is used with the series catalog. The sixth set of relationships is used to define the units. The seventh set of relationships associates laboratory methods with samples. The eighth set of relationships associates sites with the Spatial Reference system used to define the location. The last set of relationships associates project and dataset level metadata with each data source. Details of how these relationships work are given in the discussion of features of the data model design below.

Features of the Hydrologic Observations Data Model Design

Geography

The ODM is intended to be independent of the geographical representation of the site locations. Earlier versions of the data model had a shape field associated with each site. This has now been removed. The geographic location of sites is specified through the Latitude, Longitude and Elevation information in the Sites table, and optionally local coordinates which may be a standard geographic projection for the study area or a locally defined coordinate system specific to a study area. Each site also has a unique identifier, SiteID, that can be logically linked to one or more objects in a Geographic Information System (GIS) data model. For example, Figure 4 depicts a One to One relationship between sites in an Observations Data Model and HydroPoints within the Arc Hydro Framework Data Model (Maidment, 2002) used to represent objects in a digital watershed. In simple implementations, SiteID may have the same integer value as the identifier for the associated GIS object, HydroID in this case. In more general implementations it may not be possible to preserve data integrity during the merging of databases while having SiteID and HydroID the same number. In these cases where SiteID and HydroID are not the same a coupling table would be used to associate the ODM SiteIDs used to identify sites with HydroIDs. SiteID needs to be unique within ODM. This could, for example, be achieved by assigning SiteIDs from a master table. The linkage between SiteIDs and GIS object IDs is intended to be generic, suitable for use with any geographic data model that includes information specifying the location of sites. For example a linear referencing system on a river network, such as the National Hydrography Dataset, might be used to specify the location of a site on a river network. Addressing relative to specific hydrologic objects through the SiteID field provides direct and specific location information necessary for proper interpretation of values. Information from direct addressing relative to hydrologic objects is often of greater value to a user than the simple Latitude and Longitude information stored in the ODM Sites table. For example it is more useful to know that a stream gage is on such and such a stream rather than simply its latitude and longitude.

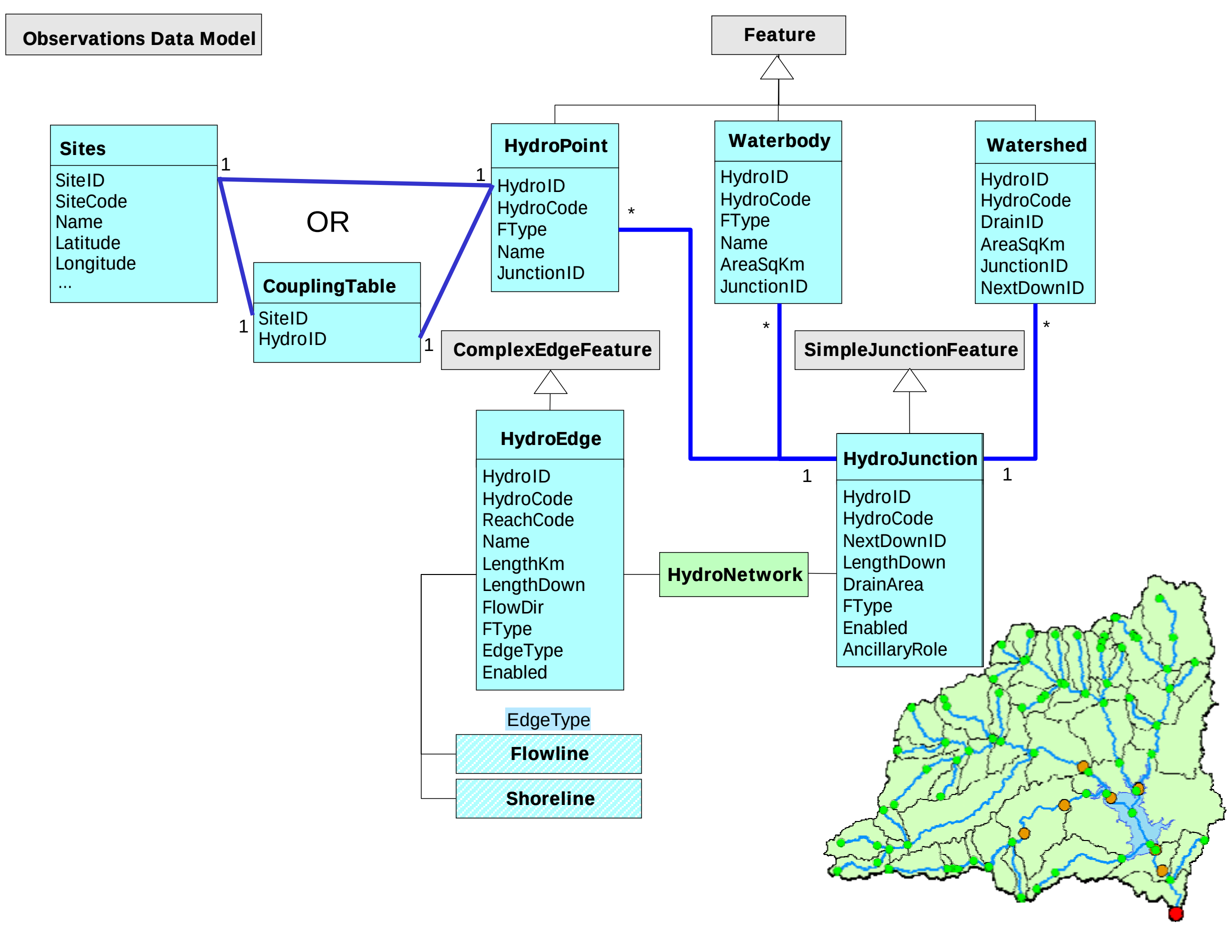


Figure 4. Arc Hydro Framework Data Model and Observations Data Model related through SiteID field in the Sites table.

Series Catalog

The SeriesCatalog table lists each unique Site, Variable combination in the Values table, and identifies each by a unique series identifier, SeriesID. This table is not required to maintain the integrity of the data but serves to provide a listing of all the distinct series of values of a specific variable at a specific site, collections of values that we refer to as data series, in the database. This entire table should be programmatically generated and modified as data is added to the database. This table provides a means by which a user can get simple descriptive information about the variables observed at a location, the most common anticipated query, without the overhead of querying the entire Values table, which can become quite large. This table is also intended to support web service queries such as GetSiteInfo that returns information about the variables available at a specific site and GetVariableInfo that returns information about the sites where a specific variable is available. The notion of data series and used here does not distinguish between different series of the same variable at the same site but measured with different offsets. So, if for example temperature was measured at two different offsets by two different sensors at one site both sets of data would fall in to one data series for the purposes of the series catalog. In these cases interpretation or analysis software will need to specifically examine and parse the offsets by examining the offset associated with each value. The series catalog does not do this, because the principal purpose of the series catalog is data discovery which we did not want to be overly complicated.

Accuracy and Precision

Each Values record has a field AccuracyStdDev. This is a numeric value intended to specify the absolute accuracy, or precision with which the value is measured, in the same units as the value. This should be specified in terms of a standard deviation (or root mean square error) of the error associated with the measurement. This is a record level attribute because it can change with the measurement protocol. For example if streamflow is measured using a V-notch weir, it is actually the stage that is measured at the precision of the depth recording instrument. The conversion to discharge through the stage-discharge relationship results in greater absolute error for larger discharges. Inclusion of this field, which will be no-data for many historic datasets because historically accuracy has not been recorded, will add to the size of data in the ODM, but will provide a way for factoring the accuracy associated with measurements into data analysis and interpretation, a trend that should be encouraged.

Offset

Each Values record has two optional fields 'Offset' and 'OffsetTypeID'. These are used to record the location of an observation relative to an appropriate datum, such as depth below the water surface, or depth below or above the ground. The OffsetTypeID references this Offset into an OffsetTypes table that gives units and definition associated with the Offset. This design only has the capability to represent one offset. In cases (which we expect to be rare) when there are multiple offsets (e.g. distance in from a stream bank and depth below the surface) one of the offsets will need to be distinguished as a separate variable.

Spatial Reference and Positional Accuracy

Unambiguous specification of the location of an observation site requires that the horizontal and vertical datum's used for latitude, longitude and elevation be specified. The spatial references table is provided for this purpose to record the name and EPSG code of each spatial reference system used. EPSG codes are numeric codes associated with coordinate system definitions published by the OGP Surveying and Positioning Committee (http://www.epsg.org/). A non-standard spatial reference system, such as for example a local grid at an experimental watershed may be defined in the spatial references table notes field. The accuracy with which the location of an observation site is known is quantified using the PosAccuracy_m field in sites. This is a numeric value intended to specify the uncertainty (as a standard deviation or root mean square error) in the spatial location information (latitude and longitude or local projection) in meters. Using a large number for PosAccuracy_m (e.g. 2000 m) accommodates entry of data collected for a study area (e.g. Panola Mountain experimental watershed) where the precise location where the observation was recorded is not known.

<u>Groups and derived from associations</u>

The DerivedFrom and Groups tables fulfill the function of grouping values for different purposes. These are tables where the same identifier (DerivedFromID or GroupID) can appear multiple times in the table associated with different ValueIDs, thereby defining the associated group of records. In the DerivedFrom table this is the sole purpose of the table and each group so defined is associated with a record in the Values table (through the DerivedFromID field in that table). This record would have been derived from the values identified by the group. The method of derivation would be given through the methods table associated with the value. This construct is useful for example to identify the 96 15 min unit streamflow values that go into the estimate of the mean daily streamflow. Note that there is no limit as to how many groups a value may be associated with, and values that are derived from other values may themselves belong to groups used to derive other values (e.g. the daily minimum flow over a month derived from daily values derived from 15 min unit values). Note also that a derived from group may have as few as one value for the case where a value is derived from a single more primitive value (e.g. Discharge from Stage). Through this construct the data model has the capability to store raw observation values and information derived from raw observations, while preserving the connection of each value to its more primitive raw measurement.

The GroupID relationship that appears in Table 2 is designated as one to many because there will be many records in Groups that have the same GroupID, but different ValueID, that serve to define the group. Similarly, there will be many records in DerivedFrom that have the same DerivedFromID, but different ValueID that serve to define the group of values from which a value is derived. Logically a value should not be in a DerivedFrom group upon which it is derived from. If this can be programmatically checked by the system, then this sort of circularity error could be prevented. The method description in the Methods table associated with a value that has a DerivedFromID should describe the method used for deriving the particular value from other values (e.g. calculating discharge from a number of velocity measurements across a stream). The relationship between the Values table DerivedFromID field and DerivedFrom table DerivedFromID field is Many to Many (*<->*) because it can occur that the same group of values is used to derive more than one derived value.

<u>Dates and times</u>

Unambiguous interpretation of date and time information requires specification of the time zone or offset from universal time (UTC). A UTCOffset field is included in the Values table to ensure that local times recorded in the database can be referenced to standard time and to enable comparison of results across databases that may store values collected in different time zones (e.g. compare values from one hydrologic observatory to those collected at another hydrologic observatory located across the country). A design choice here was to have UTCOffset as a record level qualifier because even though the time zone, and hence offset, is likely the same for all measurements at a site, the offset may change due to daylight savings. Some investigators may run data loggers on standard time, while others may adjust for daylight saving or use universal time. To avoid the necessity to keep track of the system used, or impose a system that might be cumbersome and lead to errors, we decided that if the offset was always recorded the precise time would be unambiguous and would reduce the chance for interpretation errors. A field UTCDateTime is also included as a record level attribute associated with each value. This provides a consistent time for querying and sorting data values. There is a level of redundancy between DateTime, UTCOffset and UTCDateTime. Only two are required to calculate the third. For simplicity and clarity we retain all three. A specific database implementation may choose to retain only two and calculate the third on the fly. ODM data loaders should only require two of the quantities to be input and should then calculate the third.

The separation of the date and time specification into two variables: DateTime and UTCOffset in the generic conceptual model may be handled differently within specific implementations. In one specific implementation these may be grouped in one text field in standard (e.g. ISO 8601) format such as YYYY-MM-DDhh:mm:ss.sss:UTCoffset (e.g. 2006-03-2516:19:56.232:-7), while in another format DateTime may be specified as the number of fractional days from an origin (e.g. Excel represents the above date as the following number 38801.6805 and allows the user to specify the format for display) with UTCOffset a separate variable. In general we expect specific implementations to take advantage of the representation of date time objects provided by the implementation software, but to expose the DateTime and UTCOffset to users so that time may be unambiguously interpreted.

Support scale

In interpreting values that comprise a time series it is important to know the scale information associated with the values. Blöschl and Sivapalan (1995) review the important issues. Any set of values is quantified by a scale triplet comprising support, spacing and extent, illustrated in Figure 5

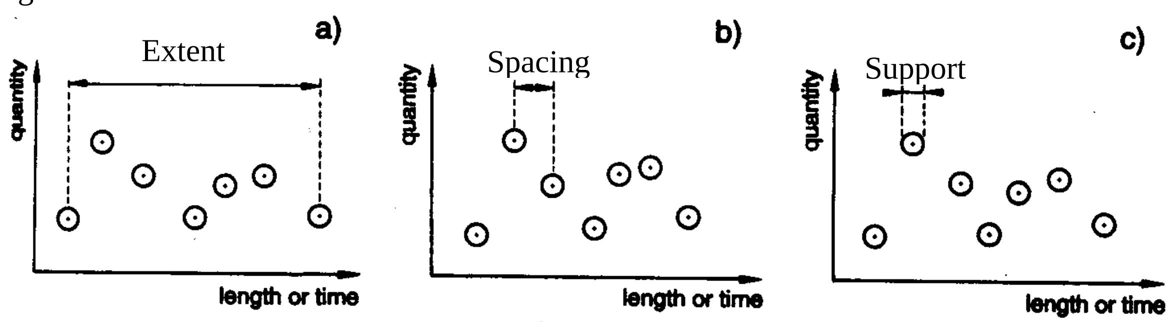


Figure 5. The Scale Triplet of Measurements (a) Extent, (b) Spacing, (c) Support. (from Blöschl, 1996)

Extent is the full range over which the measurements occur, spacing is the spacing between measurements and support is the averaging interval or footprint implicit in any measurement. In the Observations Data Model extent and spacing are properties of multiple measurements and are defined by the DateTime associated with values. We have included a field called TimeSupport in the time series table to explicitly quantify support. Figure 6 shows some of the implications associated with support, spacing and extent in the interpretation of time series values.

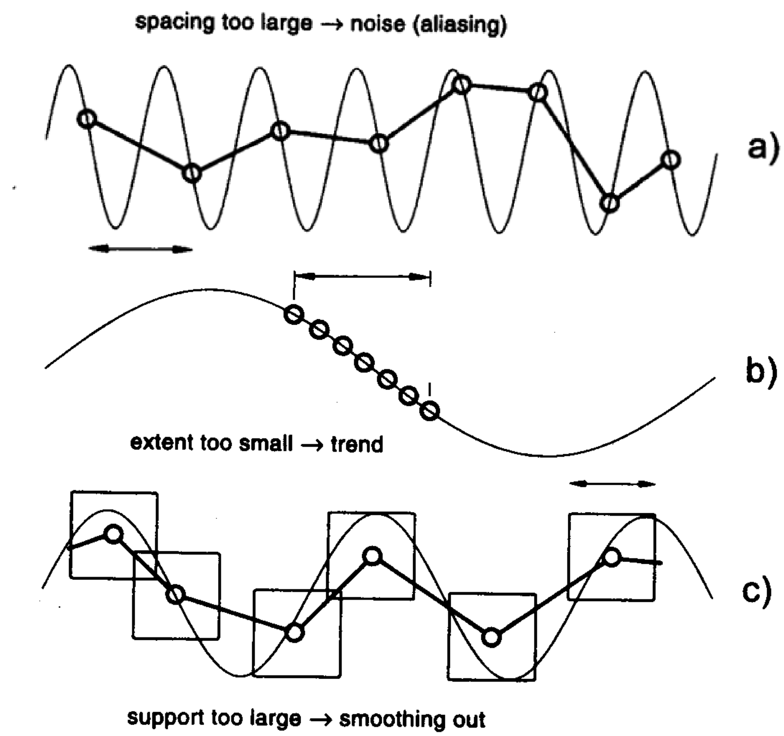


Figure 6. The effect of sampling for measurement scales not commensurate with the process scale. (a) Spacing's larger than the process scale cause aliasing in the data; (b) Extents smaller than the process scale cause a trend in the data; (c) Supports larger than the process scale cause excessive smoothing in the data. (from Blöschl, 1996)

The concepts of scale described here apply in spatial as well as time dimensions. However TimeSupport is only used to quantify support in the time dimension. The spatial support associated with a specific measurement method needs to be given or implied in the methods description in the Methods table.

Data types

In the Observations Data model the following data types are defined. These are specified by the DataType field in the Variables table.

1. *Continuous* data – the phenomenon, such as streamflow, Q(t) is specified at a particular instant in time and measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon.
2. *Instantaneous* data – the phenomenon is sampled at a particular instant in time but with a frequency that is too coarse for interpreting the record as continuous. This would be the case when the spacing is significantly larger than the support and the time scale of fluctuation of the phenomenon, such as for example infrequent water quality samples.

3. *Cumulative* data – the data represents the cumulative value of a variable measured or calculated up to a given instant of time, such as cumulative volume of flow or cumulative precipitation: $V(t) = \int_0^t Q(\tau)d\tau$, where τ represents time in the integration over the interval [0,t]. To unambiguously interpret cumulative data one needs to know the time origin. We suggest the convention of using a cumulative record with a value of zero to initialize or reset cumulative data. With this convention cumulative data should be interpreted as the accumulation over the time interval between the DateTime of the zero record and the current record at the same site position. Site position is defined by a unique combination of SiteID, VariableID, Offset and OffsetType. All four of these quantities comprise the unambiguous description of the position of an observation value and there may be multiple time series associated with multiple observation positions (e.g. redundant rain gauges with different offsets) at a location.
4. *Incremental* data – the value represents the incremental value of a variable over a time interval Δt such as the incremental volume of flow, or incremental precipitation: $\Delta V(t) = \int_{t-\Delta t}^{t} Q(\tau)d\tau$. As for cumulative data, unambiguous interpretation requires knowledge of the time increment. Here we suggest the convention of using TimeSupport if this is given, or the time interval from the previous value at the same position if TimeSupport is not given or is 0. This accommodates incremental type precipitation data that is only reported when the value is non-zero, such as NCDC data.
5. *Average* data – the value represents the average over a time interval, such as daily mean discharge or daily mean temperature: $\overline{Q}(t) = \frac{\Delta V(t)}{\Delta t}$. The averaging interval is quantified by TimeSupport in the case of regular data (as quantified by the IsRegular field) and by the time interval from the previous value at the same position for irregular data.
6. *Maximum* data – the value is the maximum value occurring at some time during a time interval, such as annual maximum discharge or a daily maximum air temperature. Again unambiguous interpretation requires knowledge of the time interval. We suggest the convention that the time interval is the TimeSupport for regular data and the time interval from the previous value at the same position for irregular data.
7. *Minimum* data – the value is the minimum value occurring at some time during a time interval, such as 7-day low flow for a year, or the daily minimum temperature. The time interval is defined similarly to Maximum data.
8. *Constant over interval* data – the value is a quantity that can be interpreted as constant over the time interval from the previous measurement.
9. *Categorical* data – the value is a categorical rather than continuous valued quantity. Mapping from Value values to categories is through the Categories table.

<u>End of interval reporting time for interval values</u>

Data types 3 to 8 above apply to values that occur over an interval of time. The DateTime reported and entered in to the data base associated with each interval value is the end time of the observation interval. This is consistent with the time a data logger would log an observation value. This is also consistent with the way the integrals in data types 3 and 4 are written, with integration up to time t.

Time Series Data

A considerable portion of hydrologic observations data is in the form of time series. This was why the initial model was based on the Arc Hydro Time Series Data Model. The ODM design has not specifically highlighted time series capabilities, nevertheless the data model has inherited the key components from the Arc Hydro Time Series Data Model to give it time series capability. In particular one variable DataType is "Continuous," designed to indicate that the values are collected with sufficient frequency as to be interpreted as a smooth time series. The IsRegular field also facilitates time series analysis because certain time series operations (e.g. Fourier Analysis) are predisposed to regularly sampled data. At first glance it may appear that there is redundancy between the IsRegular field and the DataType "Continuous" but we chose to keep these separate because there are regularly sampled quantities for which it is not reasonable to interpret the values as "Continuous". For example monthly grab samples of water quality are not continuous, but are better categorized as having DataType "Instantaneous". Note that the data model does not explicitly store the time interval between measurements, nor does it indicate where a continuous series has data gaps. Both these are required for time series analysis, but are inherently not properties of single measurements. The time interval is the time difference between sequential regular measurements, something that could be easily computed from DateTime values by analysis tools. The inference of measurement gaps (and what to do about them) from DateTime values we also regard as analysis functionality left for the Hydrologic Analysis System to handle.

Categorical Variables

In the design presented we have represented categorical or ordinal variables in the same table as continuous valued 'real' variables through a numerical encoding of the categorical value as a ‘real’ value. The Categories table then associates, for each variable, a value with an associated category description. This is a somewhat cumbersome construct because real valued quantities are being used as database keys. We do not see this as a significant shortcoming though, because typically, in our judgment, only a small fraction of hydrologic observations will be categorical. An alternative approach could have been to have a separate Values table for categorical values. The Categories table stores the categories associated with categorical values. If a Variable has a DataType that is categorical then the VariableID must match one or more VariableIDs in Categories that define the mapping between Values and Categories. The Description field in the Categories table defines the category.

Samples and Methods

At first glance there may appear to be redundancy between the information in the Samples table and Methods table. However, the samples table is intended to only be used where values are derived from a physical sample that is later analyzed in a lab, e.g. a water chemistry sample or biological sample. SampleID that links into the Samples table provides tracking of the specific physical sample used to derive each measurement and by reference to information in the LabMethods table the laboratory methods and protocols followed. The methods table refers to the method of field data collection which may specify "how" a physical sample was collected, e.g. from an automated sampler, or collected manually, but is also used to specify the measurement method associated with an in-situ measurement instrument such as a weir, turbidity probe, hydrolab, humidity or temperature sensor.

Quality Control Encoding

Each value has a field QualityControlLevel. This is an integer value between 0 and 4 intended to record the level of quality control that the value has been subjected to. The following level definitions from Earthscope (http://www.earthscope.org/links_pubs/documents/ESdatalevels_ESEC.pdf, accessed 4/9/06) are suggested so that CUAHSI ODM is consistent with the practice of other NSF Earth Observatory projects:

- **Level 0 - Raw Data**
  Raw data is defined as unprocessed data and data products that have not undergone quality control. Depending on the data type and data transmission system, raw data may be available within seconds or minutes after real-time. *Examples include real time precipitation, streamflow and water quality measurements.*

- **Level 1 – Quality Controlled Data**
  Quality controlled data have passed quality assurance procedures such as routine estimation of timing and sensor calibration or visual inspection and removal of obvious errors. *An example is USGS published streamflow records following parsing through USGS quality control procedures.*

- **Level 2 –Derived Products**
  Derived products require scientific and technical interpretation and include multiple-sensor data. *An example might be basin average precipitation derived from rain gages using an interpolation procedure.*

- **Level 3 –Interpreted Products**
  These products require researcher (PI) driven analysis and interpretation, model-based interpretation using other data and/or strong prior assumptions. *An example is basin average precipitation derived from the combination of rain gages and radar return data.*

- **Level 4 –Knowledge Products**
  These products require researcher (PI) driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions. *An example is percentages of old or new water in a hydrograph inferred from an isotope analysis.*

These definitions for quality control level are saved in the table QualityControlLevels.

Metadata

The ODM has been designed to contain all the core elements of the CUAHSI HIS metadata system (http://www.cuahsi.org/his/metadata.html) required for compliance with evolving standards such as the draft ISO 19115. The ODM in its design embodies much record, variable, and site level metadata. Dataset and project level metadata required by these standards, such as TopicCategory, Title, Abstract are included in a table ISOMetaData linked to each data source.

Reference Documents

The Methods, Sources, LabMethods and ISOMetaData tables contain links to source or reference information. At the general conceptual level of the ODM we do not specify how, or in what form these links to references or sources should be implemented. Options include using URLs or storing entire documents in the database. If external URLs are used it will be important as the database grows and is used over time to ensure that links or URL's included are stable. An alternative approach to external links is to exploit the capability of modern databases to store as fields within a record entire digital documents, such as an html or xml page, PDF document or raw data file. The capability therefore exists to instead have these links refer to a Documents table that would actually contain this metadata information, instead of housing it in digital library. There is some merit in this because then any data exported in Observations Data model format could take with it the associated metadata required to completely define it as well as the raw data upon which it is derived. This however has the disadvantage of increasing (perhaps substantially) the size of database file containing the data and being distributed to users.

Controlled Vocabularies

The following tables in the ODM are tables where controlled vocabularies for the fields are required to maintain consistency and avoid the use of synonyms that can lead to ambiguity:

- Units
- QualityControlLevel
- SpatialReferences
- SampleMediumCV
- SampleTypeCV
- CensorCodeCV
- GeneralCategoryCV
- VaraibleCV
- TopicCategoryCV
- ValueTypeCV
- VerticalDatumCV
- DataTypeCV

In the Microsoft Access database file ODMSchema.mdb provided as an electronic appendix to this document we have included initial values for the controlled vocabulary tables. We anticipate however that as experience is gained in working with the ODM that additions to the controlled vocabularies will be required. The CUAHSI HIS team welcomes input on the controlled vocabularies.

Examples

The following examples show the capability of the ODM data structure to store different types of hydrologic observations. It is not possible in examples such as this to present all the field values for all the tables so the examples present selected fields and tables chosen to illustrate key capabilities of the data model. Refer to the appendix for the complete definition of table contents.

Streamflow Gage Height and Discharge

Figure E.1 illustrates how both stream gage height measurements and the associated discharge estimates derived from the gage height measurements can be stored in the observations database. Note that gage height in feet and discharge in cubic feet per second are both in the same data table but with different VariableID that references the variables table that specifies the variable, units and other quantities associated with these values. The link between VariableID in the Values table and Variables table is shown. In this example, discharge measurements are derived from gage height (stage) measurements through a rating curve. The MethodID associated with each discharge record references into the Methods table that describes this and provides a URL that should contain metadata details for this method. The DerivedFromID in the Values table references into the DerivedFrom table that references back to the corresponding gage height in the Values table from which the discharge was derived.

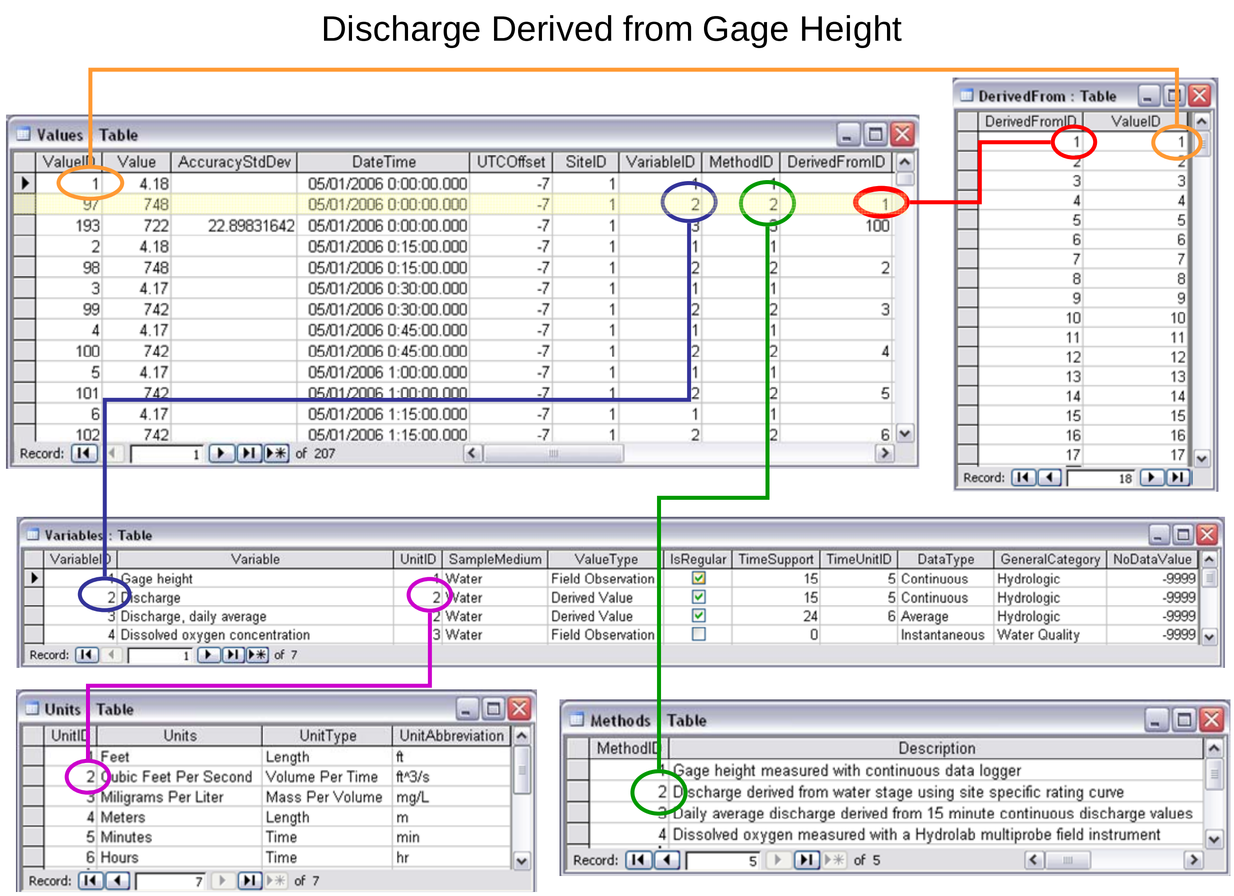
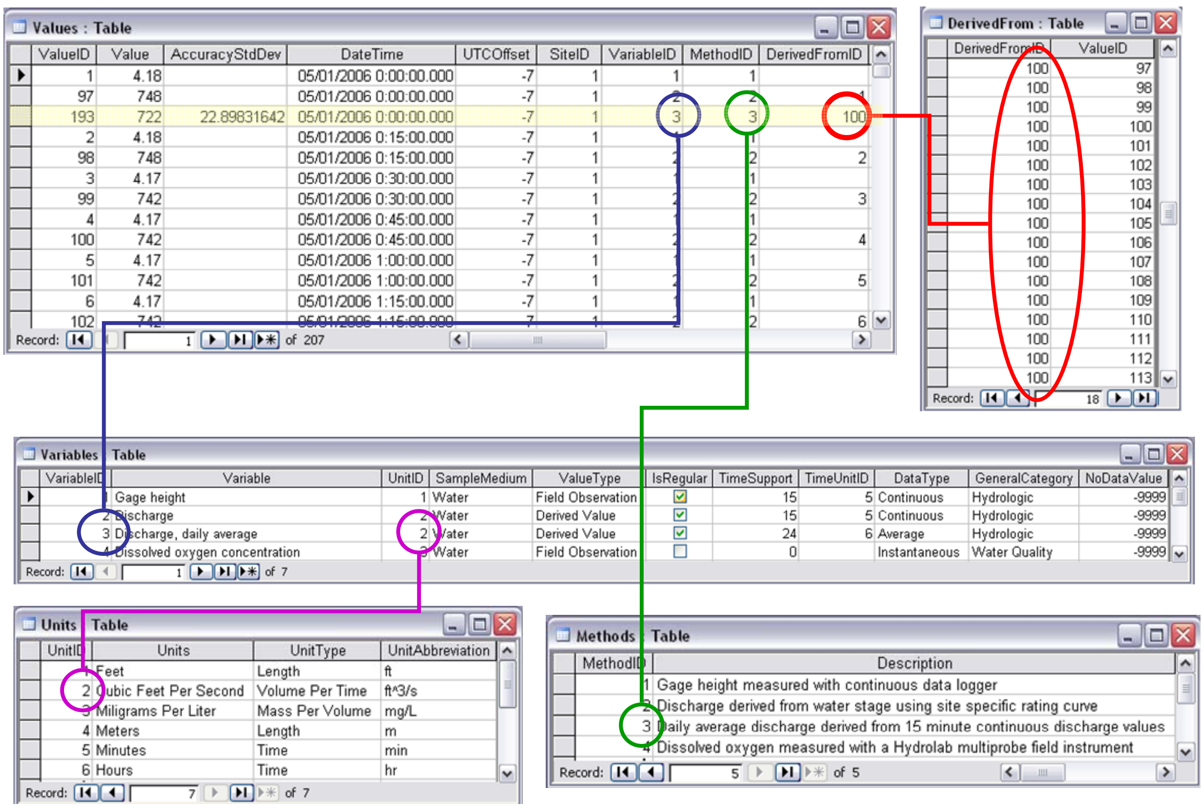


Figure E.1. Excerpts from tables illustrating the population of the data model with streamflow gage height (stage) and discharge data.

Streamflow Daily Average Discharge

Daily average streamflow is reported as an average of continuous 15 minute interval values. Figure E.2 shows excerpts from tables illustrating the population of the data model with both the continuous discharge values and derived daily averages. The record giving the single daily average discharge with Value of 722 ft^3/s in the Values table has DerivedFromID of 100. This refers to multiple records in the DerivedFrom table, with associated ValueIDs 97, 98, 99, … 113 shown. These refer to the specific 15 minute discharge values in the Values table used to derive the average daily discharge. VariableID in the Values table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of ft^3/s from UnitID referencing in to the Units table. MethodID in the Values table identifies the appropriate record in the Methods table specifying that the method used to obtain this value was daily averaging.

Daily Average Discharge Derived from 15 Minute Discharge Data



Values : Table

ValueID	Value	AccuracyStdDev	DateTime	UTCOffset	SiteID	VariableID	MethodID	DerivedFromID
1	4.18		05/01/2006 0:00:00.000	-7	1	1	1	
97	748		05/01/2006 0:00:00.000	-7	1	2	2	1
193	722	22.89831642	05/01/2006 0:00:00.000	-7	1	3	3	100
2	4.18		05/01/2006 0:15:00.000	-7	1	1	1	
98	748		05/01/2006 0:15:00.000	-7	1	2	2	2
3	4.17		05/01/2006 0:30:00.000	-7	1	1	1	
99	742		05/01/2006 0:30:00.000	-7	1	2	2	3
4	4.17		05/01/2006 0:45:00.000	-7	1	1	1	
100	742		05/01/2006 0:45:00.000	-7	1	2	2	4
5	4.17		05/01/2006 1:00:00.000	-7	1	1	1	
101	742		05/01/2006 1:00:00.000	-7	1	2	2	5
6	4.17		05/01/2006 1:15:00.000	-7	1	1	1	
102	742		05/01/2006 1:15:00.000	-7	1	2	2	6

Record: 1 of 207

DerivedFrom : Table

DerivedFromID	ValueID
100	97
100	98
100	99
100	100
100	101
100	102
100	103
100	104
100	105
100	106
100	107
100	108
100	109
100	110
100	111
100	112
100	113

Record: 18

Variables : Table

VariableID	Variable	UnitID	SampleMedium	ValueType	IsRegular	TimeSupport	TimeUnitID	DataType	GeneralCategory	NoDataValue
1	Gage height	1	Water	Field Observation	☑	15	5	Continuous	Hydrologic	-9999
2	Discharge	2	Water	Derived Value	☑	15	5	Continuous	Hydrologic	-9999
3	Discharge, daily average	2	Water	Derived Value	☑	24	6	Average	Hydrologic	-9999
4	Dissolved oxygen concentration	[illegible]	Water	Field Observation	☐	0		Instantaneous	Water Quality	-9999

Record: 1 of 7

Units : Table

UnitID	Units	UnitType	UnitAbbreviation
1	Feet	Length	ft
2	Cubic Feet Per Second	Volume Per Time	ft^3/s
3	Miligrams Per Liter	Mass Per Volume	mg/L
4	Meters	Length	m
5	Minutes	Time	min
6	Hours	Time	hr

Record: 7 of 7

Methods : Table

MethodID	Description
1	Gage height measured with continuous data logger
2	Discharge derived from water stage using site specific rating curve
3	Daily average discharge derived from 15 minute continuous discharge values
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument

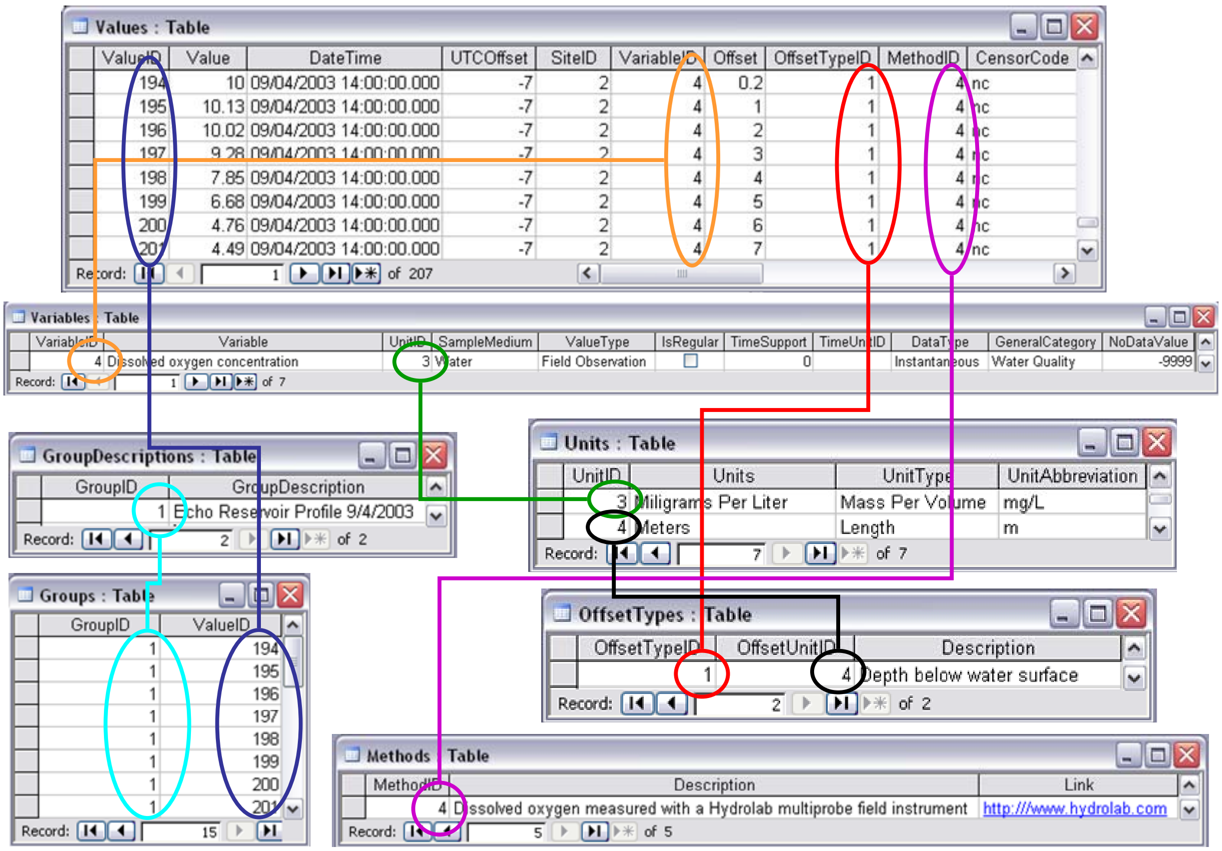
Record: 5 of 5

Figure E.2. Excerpts from tables illustrating the population of the data model with daily average discharge derived from 15 minute discharge values.

<u>Water Chemistry from a Profile in a Lake</u>

Reservoir profile measurements provide an example of the logical grouping of values, and values that have an offset in relationship to the location of the sampling station. These measurements may be made simultaneously (by multiple instruments in the water column) or over a short time period (one instrument that is lowered from top to bottom). Figure E.3 shows an example of how these data would be stored in the database structure. The OffsetTypes table and Offset attribute is used to quantify the depth associated with each measurement. Each of the values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType that the offset is Depth below water surface. The OffsetTypes table references in to the Units table indicating that the OffsetUnits are m, so Offset in the Values table is in units of m depth below the water surface. Each of the values shown also has a VariableID that in the Variables table indicates that the variable measured was dissolved oxygen concentration in units of mg/L. Each of the values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The values shown are part of a logical group of values representing the water chemistry profile in a lake. This is represented using the Groups table and GroupDescriptions table. The Groups table associates GroupID 1 with each of the ValueIDs of the values belonging to the group. A description of this group is given in the GroupDescriptions table.

Water Chemistry From a Lake Profile



Values : Table

ValueID	Value	DateTime	UTCOffset	SiteID	VariableID	Offset	OffsetTypeID	MethodID	CensorCode
194	10	09/04/2003 14:00:00.000	-7	2	4	0.2	1	4	nc
195	10.13	09/04/2003 14:00:00.000	-7	2	4	1	1	4	nc
196	10.02	09/04/2003 14:00:00.000	-7	2	4	2	1	4	nc
197	9.28	09/04/2003 14:00:00.000	-7	2	4	3	1	4	nc
198	7.85	09/04/2003 14:00:00.000	-7	2	4	4	1	4	nc
199	6.68	09/04/2003 14:00:00.000	-7	2	4	5	1	4	nc
200	4.76	09/04/2003 14:00:00.000	-7	2	4	6	1	4	nc
201	4.49	09/04/2003 14:00:00.000	-7	2	4	7	1	4	nc

Record: 1 of 207

Variables : Table

VariableID	Variable	UnitID	SampleMedium	ValueType	IsRegular	TimeSupport	TimeUnitID	DataType	GeneralCategory	NoDataValue
4	Dissolved oxygen concentration	3	Water	Field Observation	☐	0		Instantaneous	Water Quality	-9999

Record: 1 of 7

GroupDescriptions : Table

GroupID	GroupDescription
1	Echo Reservoir Profile 9/4/2003

Record: 2 of 2

Units : Table

UnitID	Units	UnitType	UnitAbbreviation
3	Miligrams Per Liter	Mass Per Volume	mg/L
4	Meters	Length	m

Record: 7 of 7

Groups : Table

GroupID	ValueID
1	194
1	195
1	196
1	197
1	198
1	199
1	200
1	201

Record: 15

OffsetTypes : Table

OffsetTypeID	OffsetUnitID	Description
1	4	Depth below water surface

Record: 2 of 2

Methods : Table

MethodID	Description	Link
4	Dissolved oxygen measured with a Hydrolab multiprobe field instrument	http://www.hydrolab.com

Record: 5 of 5

Figure E.3. Excerpts from tables illustrating the population of the data model with Water Chemistry data.

NCDC Precipitation Data

Figure E.4 illustrates the representation of NCDC 15 min precipitation data by the Data Model. The data includes 15 min incremental values as well as daily totals. Separate records in the

Variables table are used for the 15 min or daily total values. This data is reported at irregular intervals and only logged for time periods for which precipitation is non zero. This is accommodated by setting the IsRegular attribute associated with the variable to False and specifying the TimeSupport value as 15 or 24 and the TimeUnit as "Minute" or "Hour". The DataType of "Incremental" is used to indicate that these are incremental values defined over the TimeSupport interval. The convention for incremental data (see above) is that when the time support is specified, it specifies the increment for irregular incremental data. When time support is specified as 0 it means the increment is from the previous value at the same site position. Data qualifier codes indicate periods where the data is missing. The method associated with each precipitation variable documents the convention that zero precipitation periods are not logged in this data acquired from NCDC. A data qualifier code is also used to flag days where the precipitation total is incomplete due to the record being missing during part of the day.

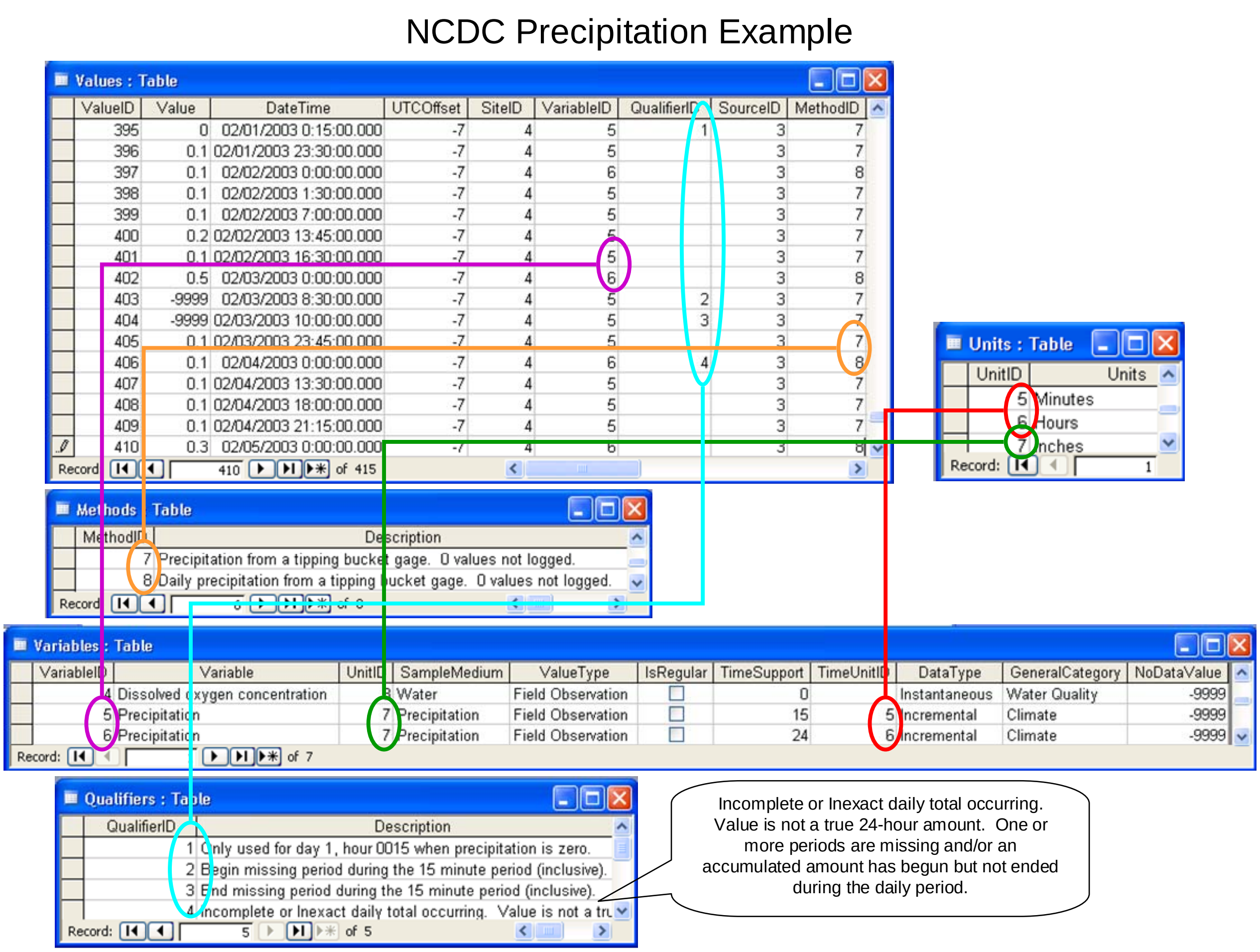
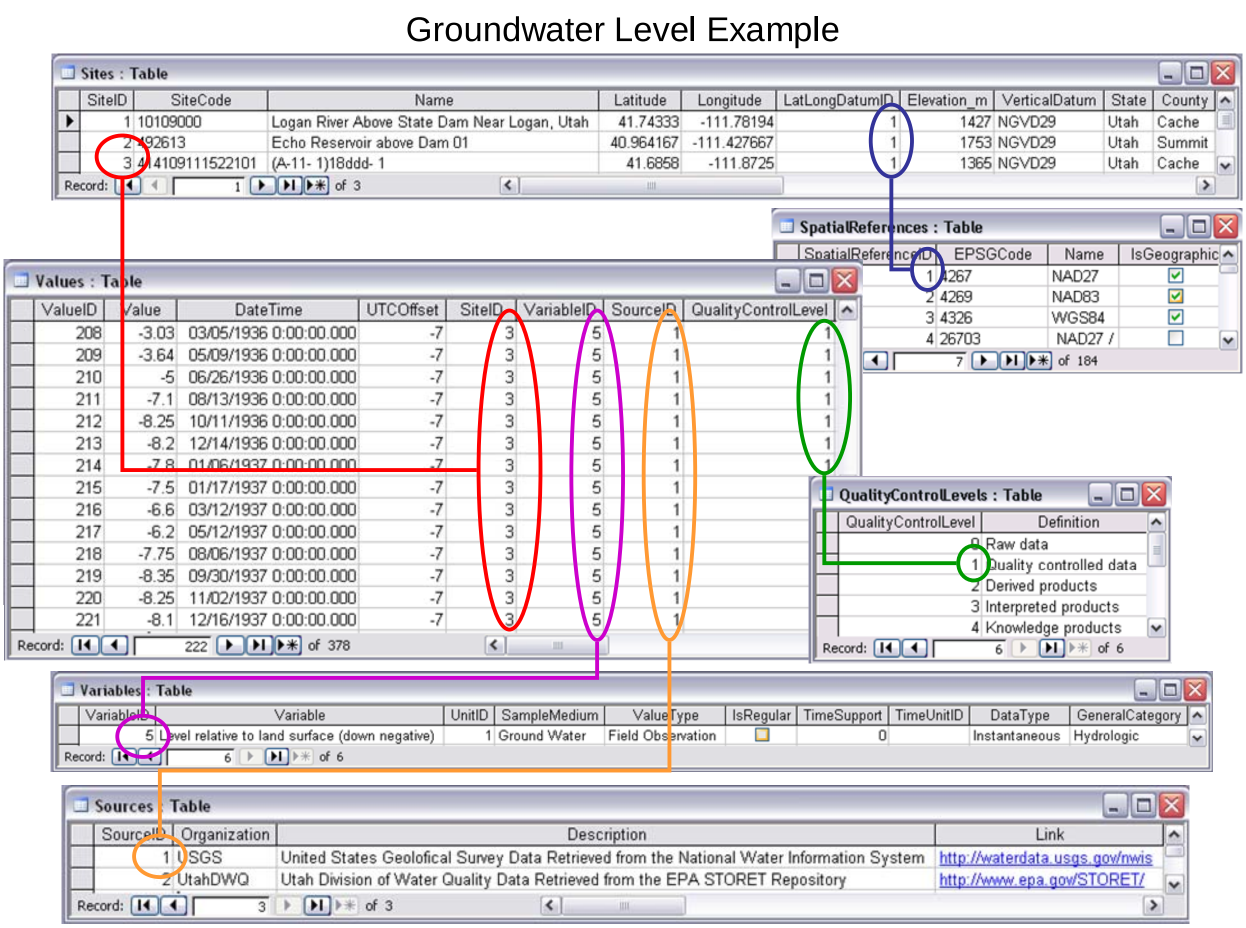


Figure E.4. Excerpts from tables illustrating the population of the data model with NCDC Precipitation Data.

Groundwater Level

The following is an example of how groundwater level data can be stored in the ODM database structure. In this groundwater level example, values are the water table level relative to the ground surface reported as negative values. This example shows multiple values of a single variable at a single site made by a single source that have been quality controlled as indicated by the QualityControlLevel field in the QualityControlLevels table. The SiteID field in the Values table indicates the site in the Sites table that gives the location information about the site. In this case the elevation is with respect to the NGVD29 datum as indicated in the VerticalDatum field and latitude and longitude are with respect to the NAD27 datum as indicated in the SpatialReferences table. The VariableID field in the Values table references the appropriate record in the Variables table indicating information about the variable. The SourceID field in the Values table references the appropriate record in the Sources table giving information about the source of the data.



SiteID	SiteCode	Name	Latitude	Longitude	LatLongDatumID	Elevation_m	VerticalDatum	State	County
1	10109000	Logan River Above State Dam Near Logan, Utah	41.74333	-111.78194	1	1427	NGVD29	Utah	Cache
2	492613	Echo Reservoir above Dam 01	40.964167	-111.427667	1	1753	NGVD29	Utah	Summit
3	414109111522101	(A-11- 1)18ddd- 1	41.6858	-111.8725	1	1365	NGVD29	Utah	Cache

SpatialReferenceID	EPSGCode	Name	IsGeographic
1	4267	NAD27	☑
2	4269	NAD83	☑
3	4326	WGS84	☑
4	26703	NAD27 /	☐

ValueID	Value	DateTime	UTCOffset	SiteID	VariableID	SourceID	QualityControlLevel
208	-3.03	03/05/1936 0:00:00.000	-7	3	5	1	1
209	-3.64	05/09/1936 0:00:00.000	-7	3	5	1	1
210	-5	06/26/1936 0:00:00.000	-7	3	5	1	1
211	-7.1	08/13/1936 0:00:00.000	-7	3	5	1	1
212	-8.25	10/11/1936 0:00:00.000	-7	3	5	1	1
213	-8.2	12/14/1936 0:00:00.000	-7	3	5	1	1
214	-7.8	01/06/1937 0:00:00.000	-7	3	5	1	1
215	-7.5	01/17/1937 0:00:00.000	-7	3	5	1	
216	-6.6	03/12/1937 0:00:00.000	-7	3	5	1	
217	-6.2	05/12/1937 0:00:00.000	-7	3	5	1	
218	-7.75	08/06/1937 0:00:00.000	-7	3	5	1	
219	-8.35	09/30/1937 0:00:00.000	-7	3	5	1	
220	-8.25	11/02/1937 0:00:00.000	-7	3	5	1	
221	-8.1	12/16/1937 0:00:00.000	-7	3	5	1	

QualityControlLevel	Definition
0	Raw data
1	Quality controlled data
2	Derived products
3	Interpreted products
4	Knowledge products

VariableID	Variable	UnitID	SampleMedium	ValueType	IsRegular	TimeSupport	TimeUnitID	DataType	GeneralCategory
5	Level relative to land surface (down negative)	1	Ground Water	Field Observation	☐	0		Instantaneous	Hydrologic

SourceID	Organization	Description	Link
1	USGS	United States Geolofical Survey Data Retrieved from the National Water Information System	http://waterdata.usgs.gov/nwis
2	UtahDWQ	Utah Division of Water Quality Data Retrieved from the EPA STORET Repository	http://www.epa.gov/STORET/

Figure E.5. Excerpts from tables illustrating the population of the data model with irregularly sampled groundwater data.

Appendix A. Observations Data Model Table and Field Structure

The following is a description of the tables in the observations data model, a listing of the fields contained in each table, a description of the data contained in each field and its type, examples of the information to be stored in each field where appropriate, and any additional information about each field. Values in the example column should not be considered to be inclusive of all potential values, especially in the case of fields that require a controlled vocabulary. Initial values for controlled vocabularies are included in the Microsoft Access database file ODMSchema.mdb provided as an electronic appendix to this document. We anticipate that these controlled vocabularies will need to be extended and adjusted. Tables appear in alphabetical order

The Validation column for each field indicates the following
A – Automatically provided by database
M – Mandatory
O – Optional
P – Programmatically derived

Table: Categories

This table defines the categories for categorical variables. Records are required for variables where DataType is "Categorical". Multiple entries for each VariableID, with different Values provide the mapping from Value to category description.

Field Name	Data Type	Description	Examples	Validation
VariableID	Integer	Integer identifier that references the Variables record of a categorical variable.	45	M
Value	Real	Numeric value	1.0	M
Description	Text	Definition of categorical variable value	"Cloudy"	M

Table: CensorCodeCV

Table that contains controlled vocabulary for censor codes.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for CensorCode	lt, gt, nc	M
Definition	Text	Definition of CensorCode controlled vocabulary. The definition is optional if the term is self explanatory.	less than, greater than, not censored.	O

Table: DataTypeCV

Table that contains the controlled vocabulary for data types.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for data type	Continuous	M
Definition	Text	Definition of data type controlled vocabulary. The definition is optional if the term is self explanatory.	A quantity specified at a particular instant in time measured with sufficient frequency (small spacing) to be interpreted as a continuous record of the phenomenon.	O

Table: DerivedFrom

Table that contains the linkage between derived quantities and the values that they were derived from.

Field Name	Data Type	Description	Examples	Validation
DerivedFromID	Integer	Integer identifying the group of values from which a quantity is derived	3	M
ValueID	Integer	Integer identifier referencing values that comprise a group from which a quantity is derived. This corresponds to ValueID in the Values table.	1,2,3,4,5	M

Table: GeneralCategoryCV

Table that contains controlled vocabulary for the general categories associated with Variables.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for GeneralCategory	Hydrology	M
Definition	Text	Definition of GeneralCategory controlled vocabulary. The definition is optional if the term is self explanatory.	Data associated with hydrologic variables or processes.	O

Table: GroupDescriptions

Lists the descriptions for each of the groups that have been formed.

Field Name	Data Type	Description	Example	Validation
GroupID	Integer	Integer identifier for each group of values that has been formed. This also references to GroupID in the Groups table.	4	M
Description	Text	Text description of the group	"Echo Reservoir Profile 7/7/2005"	O

Table: Groups

Lists the groups of values that have been created and the values that are within each group.

Field Name	Data Type	Description	Example	Validation
GroupID	Integer	Integer ID for each group of values that has been formed.	4	M
ValueID	Integer	Integer identifier for each value that belongs to a group. This corresponds to ValueID in the Values table	2,3,4	M

Table: ISOMetadata

Dataset and project level metadata required by the CUAHSI HIS metadata system (http://www.cuahsi.org/his/metadata.html) for compliance with standards such as the draft ISO 19115 or ISO 8601.

Field Name	Data Type	Description	Example	Validation
MetaDataID	Integer	Unique integer ID for each metadata record.	4	M
TopicCategory	Text	Topic category keyword that gives the broad ISO19115 metadata topic category for data from this source. The controlled vocabulary of topic category keywords is given in the TopicCategoryCV table.	inlandWaters	M
Title	Text	Title of data from a specific data source		M
Abstract	Text	Abstract of data from a specific data source		M
ProfileVersion	Text	Name of metadata profile used by the data source	ISO8601	M
Link	Hyperlink	Link to additional metadata reference material		O

Table: LabMethods

Gives descriptions of the laboratory methods used to process measurements based on physical samples.

Field Name	Data Type	Description	Example	Validation
LabMethodID	Integer	Unique integer identifier for each laboratory method. This is the key used by the Samples table to reference a method	6	M
LabName	Text	Name of the Laboratory responsible for processing the sample	"USGS Atlanta Field Office"	M
LabMethodName	Text	Name of the method and protocols used for sample analysis	USEPA-365.1	M
LabOrganization	Text	Organization responsible for sample analysis	USGS	M
Description	Text	Description of the method and protocols used for sample analysis	Processed through Model *** Mass Spectrometer	M
Link	Hyperlink	Link to additional reference material on the analysis method		O

Table: Methods

Lists the methods used to collect the data and provides an indication of the Quality Assurance and Quality Control procedures associated with each method.

Field Name	Data Type	Description	Example	Validation
MethodID	Integer	Unique integer ID for each method	5	M
Description	Text	Text description of each method including Quality Assurance and Quality Control procedures.	"Specific conductance measured using a Hydrolab" or "Streamflow measured using a V notch weir with dimensions xxx"	M
Link	Hyperlink	Link to additional reference material on the method.		O

Table: OffsetTypes

Lists full descriptive information for each of the measurement offsets.

Field Name	Data Type	Description	Example	Validation
OffsetTypeID	Integer	Unique integer identifier that identifies the type of measurement offset	2	M
OffsetUnitID	Integer	Integer identifier that references the record in the Units table giving the Units of the Offset	1	M

Description	Text	Full text description of the offset type	"Below water surface" "Above Ground Level"	M

Table: Qualifiers

Gives data qualifying comments that accompany the data.

Field Name	Data Type	Description	Example	Validation
QualifierID	Integer	Unique Integer identifying the data qualifier code	3	M
Description	Text	Text of the data qualifying comment	"Holding time for sample analysis exceeded"	M

Table: QualityControlLevels

Gives the controlled vocabulary for quality control levels.

Field Name	Data Type	Description	Example	Validation
QualityControlLevel	Integer	Integer between 0 and 4 inclusively giving the level of quality control that the value has been subjected to.	0, 1, 2, 3, 4	M
Definition	Text	Definition of Quality Control Level.	Raw Data, Quality Controlled data	M
Explanation	Text	Explanation of Quality Control Level	Raw data is defined as unprocessed data and data products that have not undergone quality control.	

Table: SampleMediumCV

Table that contains the controlled vocabulary for sample media.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for sample media	Surface Water	M
Definition	Text	Definition of sample media controlled vocabulary. The definition is optional if the term is self explanatory.	Sample taken from surface water such as a stream, river, lake, pond, reservoir, ocean, etc.	O

Table: Samples

Gives information about physical samples analyzed in a lab.

Field Name	Data Type	Description	Example	Validation
SampleID	Integer	Unique integer identifier that identifies each physical sample	3	M
SampleType	Text	Controlled vocabulary specifying the sample type	"FD", "PB", "SW"	M
LabSampleCode	Text	Code or label used to identify and track lab sample or sample container (e.g. bottle) during lab analysis.	AB-123	M
LabMethodID	Integer	Unique identifier for the laboratory method used to process sample. This references in to the LabMethods table.	4	M

Table: SampleTypeCV

Table that contains the controlled vocabulary for sample type.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for sample type	"FD", "PB"	M
Definition	Text	Definition of sample type controlled vocabulary. The definition is optional if the term is self explanatory.	"Foliage Digestion", "Precipitation Bulk"	O

Table: SeriesCatalog

Lists each separate data series in the data base for the purposes of identifying or displaying what data are available at each location and to speed simple queries without querying the main Values table. Unique site – variable combinations are defined by unique combinations of SiteID and VariableID.

This entire table should be programmatically derived.

Field Name	Data Type	Description	Example	Validation
SeriesID	Integer	Unique integer identifier for each data series	5	P/M
SiteID	Integer	Site identifier from Sites table.	7	P/M
SiteCode	Text	Site identifier used by organization that collects the data.	1002000	P/O
Name	Text	Full text name of sampling location	"Logan River"	P/O
VariableID	Integer	Integer identifier for each Variable that references the Variables table.	4	P/M
Variable	Text	Name of the variable from the variables table	Temperature	P/M
UnitID	Integer	Integer identifier that references the record in the Units table giving the Units of the value	5	M

Field Name	Data Type	Description	Example	Validation
SampleMedium	Text	The medium of the sample. This should be from the SampleMediumCV controlled vocabulary table.	Water	P/M
ValueType	Text	Text value indicating what type of value is being recorded. This should be from the ValueTypeCV controlled vocabulary table.	Field Observation	P/M
BeginDateTime	Date/Time	Date of the first value in the series. To be programmatically updated if new records are added.	9/4/2003 7:00:00 AM	P
EndDateTime	Date/Time	Date of the last value in the series. To be programmatically updated if new records are added.	9/4/2005 7:00:00 AM	P
UTCOffset	Integer	Offset from UTC time at the sampling location.	-7	Number of hours
ValueCount	Integer	The number of values in the series identified by the combination of the SiteID and the VariableID	50	P (programmatically updated if new records are added)
GeneralCategory	Text	General category of the variable from the GeneralCategoryCV table.	Water Quality	

Table: Sites

Provides information giving the spatial location at which values have been collected.

Field Name	Data Type	Description	Example	Validation
SiteID	Integer	Unique identifier for each sampling location.	37	M
SiteCode	Text	Site identifier used by organization that collects the data	“10109000” (USGS Gage number)	O
Name	Text	Full name of sampling location	“LOGAN RIVER ABOVE STATE DAM, NEAR LOGAN,UT”	O
Latitude	Real	Latitude in decimal degrees	45.32	M
Longitude	Real	Longitude in decimal degrees. E positive, W negative	-100.47	M
LatLongDatumID	Integer	Identifier that references spatial reference system in SpatialReferences table	1	M
Elevation_m	Real	Elevation of sampling location (in m). If this is not provided it needs to be obtained programmatically from a DEM based on location information.	1432	M/P
VerticalDatumID	Text	Vertical Datum. Controlled Vocabulary from VerticalDatumCV	NAVD88	M
LocalX	Real	Local Projection X coordinate	456700	O
LocalY	Real	Local Projection Y Coordinate	232000	O

Field Name	Data Type	Description	Example	Validation
LocalProjectionInfo	Integer	Identifier that references spatial reference system in SpatialReferences table	7	O
PosAccuracy_m	Real	Value giving the accuracy with which the positional information is specified in meters	100	O
State	Text	Name of state in which the sampling station is located	"Utah"	O
County	Text	Name of County in which the sampling station is located	"Cache"	O
Comments	Text	Comments related to the site		O

Table: Sources

Lists the original sources of the data, providing information sufficient to retrieve and reconstruct the value from the original data files if necessary.

Field Name	Data Type	Description	Example	Validation
SourceID	Integer	Unique integer identifier that identifies each data source	5	M
Organization	Text	Name of Organization that collected the data. This should be the agency or organization that collected the data, even if it came out of a database consolidated from many sources such as STORET.	Utah Division of Water Quality	M
Description	Text	Full text description of the source of the data	"Text file retrieved from the EPA STORET system indicating data originally from Utah Division of Water Quality"	M
Link	Hyperlink	Link to original data file and associated metadata stored in the digital library or URL of data source		M
ContactName	Text	Name of Contact Person for data source	Jane Adams	M
Email	Text	Email address for contact person	Jane.Adams@dwq.u t	M
Address	Text	Address for contact person	45 Main Street	M
City	Text	City for contact person	Salt Lake City	M
State	Text	State for Contact Person. Use two letter abbreviations for US. For other countries give full country name	UT, New Zealand	M
ZipCode	Text	US Zip Code or country postal code	82323	M
MetaDataID	Integer	Integer identifier referencing the record in the ISOMetaData table for this source	5	M

Table: SpatialReferences

Provides information about the spatial reference systems used for latitude and longitude as well as local coordinate systems in Sites.

Field Name	Data Type	Description	Example	Validation
SpatialReferenceID	Integer	Unique identifier for each Spatial Reference System	37	M
EPSGCode	Text	Code for Spatial Reference System from http://www.epsg.org/	4269	O
Name	Text	Name of Spatial Reference System	NAD83	O
IsGeographic	Boolean	Value that indicates whether the spatial reference system uses geographic coordinates (i.e. latitude and longitude) or not	"True", "False"	M
Notes	Text	Descriptive information about spatial reference system. This field would be used to define a non-standard study area specific system if necessary. Description of local projection information. Where possible this should refer to a standard projection in which case latitude and longitude can be determined from local projection information. If the local grid system is non-standard then latitude and longitude need to be included too.		O

Table: TopicCategoryCV

Table that contains controlled vocabulary for the ISOMetaData topic categories.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for TopicCategory	InlandWaters	M
Definition	Text	Definition of TopicCategory controlled vocabulary. The definition is optional if the term is self explanatory.	Data associated with inland waters	O

Table: Units

Gives the Units and UnitType associated with observations. This is a controlled vocabulary table.

Field Name	Data Type	Description	Example	Validation
UnitID	Integer	Unique integer identifier that identifies each unit	6	M
Units	Text	Text units of the value.	“Degrees Celsius”	M
UnitType	Text	Text value that specifies the dimensions of the units	“Length” “Time” “Mass”	M
UnitAbbreviation	Text	Abbreviation for Units	C	M

Table: Values

Stores the actual data values.

Field Name	Data Type	Description	Example	Validation
ValueID	Integer	Unique integer identifier for each value	43	M
Value	Real	The numeric value. For Categorical variables a number is stored here. The Variables table has DataType as Categorical and the Categories table maps from number on to Category Description.	34.5	M
AccuracyStdDev	Real	Numeric value that describes the measurement accuracy of the value, expressed as the estimated standard deviation (or root mean square error) of the specific value. If not given it is interpreted as unknown.	4	O
DateTime	Date/Time	Local date and time at which the value was observed. Represented in an implementation specific format	9/4/2003 7:00:00 AM	M
UTCOffset	Real	Offset in hours from UTC time at the sampling location	-7	M
UTCDateTime	Date/Time	Universal UTC date and time at which the value was observed. Represented in an implementation specific format	9/4/2003 2:00:00 PM	M
SiteID	Integer	Unique identifier for each sampling location. This links values to their locations in the Sites table	3	M
VariableID	Integer	Integer identifier that references the variable that was measured. This links values to their type in the Variables table.	5	M
Offset	Real	Distance from a datum or control point at which a value was observed. If not given the offset is inferred to be 0, or not relevant/necessary.	2.1	O
OffsetTypeID	Integer	Identifier that references the measurement offset type in the OffsetTypes table.	3	O
CensorCode	Text	Text indication of whether the data value is censored, from the CensorCodeCV controlled vocabulary. If null or not given the value is inferred to not be censored.	nc	O
QualifierID	Integer	Integer identifier that references the Qualifiers table. If Null the value is inferred to not be qualified	4	O
MethodID	Integer	Integer identifier that references the Methods table.	3	M
SourceID	Integer	Integer identifier that references the record in the Sources table giving the source of the value.	5	M
SampleID	Integer	Integer Identifier that references into the Samples table. This is required only if the value is from a physical sample processed in a lab	7	O

Field Name	Data Type	Description	Example	Validation
DerivedFromID	Integer	Integer identifier for the derived from group of values that the current value is derived from. This refers to a group of derived from records in the DerivedFrom table. If null or not given the value is inferred to not be derived from another value.	5	O
QualityControlLevel	Integer	Integer between 0 and 4 inclusively giving the level of quality control that the value has been subjected to. This references in to QualityControlLevels table.	1	O

Table: ValueTypeCV

Table that contains controlled vocabulary for the ValueType field in Variables and SeriesCatalog.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for ValueType	Field Observation	M
Definition	Text	Definition of ValueType controlled vocabulary. The definition is optional if the term is self explanatory.	Observation of a variable using a field instrument	O

Table: VariableCV

Table that contains controlled vocabulary for the Variable field in Variables and SeriesCatalog.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for Variable	"Temperature", "Discharge", "Precipitation"	M
Definition	Text	Definition of Variable controlled vocabulary. The definition is optional if the term is self explanatory.		O

Table: Variables

Lists the full descriptive information about what variables have been measured.

Field Name	Data Type	Description	Example	Validation
VariableID	Integer	Unique integer identifier for each Variable	6	M
Variable	Text	Name of the variable that was measured, observed, modeled, etc. This should be from the VariableCV controlled vocabulary table.	“Discharge”	M
UnitID	Integer	Integer identifier that references the record in the Units table giving the Units of the value	4	M

Field Name	Data Type	Description	Example	Validation
SampleMedium	Text	The medium of the sample. This should be from the SampleMediumCV controlled vocabulary table.	“Surface Water” “Sediment” “Fish Tissue”	M
ValueType	Text	Text value indicating what type of value is being recorded. This should be from the ValueTypeCV controlled vocabulary table.	“Field Observation” “Laboratory Observation” “Model Simulation Results”	M
IsRegular	Boolean	Value that indicates whether the values are from a regularly sampled time series	“True” “False”	M
TimeSupport	Real	Numerical value that indicates the time support (or temporal footprint). 0 is used to indicate a value that is instantaneous. Other values indicate the time over which the values are implicitly or explicitly averaged.	0, 24	M
TimeUnitID	Text	Integer identifier that references the record in the Units table giving the Units of the time support. If TimeSupport is 0, indicating an instantaneous observation, a unit needs to still be given for completeness, although it is somewhat arbitrary.	4	M
DataType	Text	Text value that identifies the data as one of several types from the DataType CV controlled vocabulary table.	“Continuous” “Instantaneous” “Cumulative” “Incremental” “Average” “Minimum: “Maximum” “Constant Over Interval” “Categorical”	M
GeneralCategory	Text	General category of the values from the GeneralCategoryCV controlled vocabulary table.	“Climate” “Water Quality” “Groundwater Quality”	M
NoDataValue	Real	Numeric value used to encode no data values for this variable.	-9999	M

Table: VerticalDatumCV

Table that contains controlled vocabulary for the VerticalDatum field in Sites.

Field Name	Data Type	Description	Examples	Validation
Term	Text	Controlled vocabulary for VerticalDatum	NAVD88	M
Definition	Text	Definition of VerticalDatum controlled vocabulary. The definition is optional if the term is self explanatory.	North American Vertical Datum of 1988	O

References

Blöschl, G., (1996), Scale and Scaling in Hydrology, Habilitationsschrift, Weiner Mitteilungen Wasser Abwasser Gewasser, Wien, 346 p.

Blöschl, G. and M. Sivapalan, (1995), "Scale Issues in Hydrological Modelling: A Review," Hydrological Processes, 9(1995): 251-290.

Horsburgh, J. S., D. G. Tarboton and D. R. Maidment, (2005), "A Community Data Model for Hydrologic Observations, Chapter 6," in Hydrologic Information System Status Report, Version 1, Edited by D. R. Maidment, p.102-135, http://www.cuahsi.org/docs/HISStatusSept15.pdf.

Maidment, D. R., ed. (2002), Arc Hydro Gis for Water Resources, ESRI Press, Redlands, CA, 203 p.

Maidment, D. R., (2005a), "A Data Model for Hydrologic Observations." Paper prepared for presentation at the CUAHSI Hydrologic Information Systems Symposium, University of Texas at Austin. March 7, 2005.

Maidment, D. R., ed. (2005b), 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.

Tarboton, D. G., (2005), "Review of Proposed Cuahsi Hydrologic Information System Hydrologic Observations Data Model." Utah State University. May 5, 2005.

Tarboton, D. G., J. S. Horsburgh and D. R. Maidment, (2006), "Cuahsi Community Hydrologic Observations Data Model Working Design Specifications Document - Version 2," http://www.cuahsi.org/his/tk-observedb.html.