CUAHSI Community Observations Data Model (ODM) Version 1.0 Design Specifications

March 7, 2007

David G. Tarboton[1], Jeffery S. Horsburgh[1], David R. Maidment[2]

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 ODM 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. Although designed specifically with hydrologic observation data in mind, this data model has a simple and general structure that will also

[1] Utah Water Research Laboratory, Utah State University
[2] Center for Research in Water Resources, University of Texas at Austin

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, 2005) 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. Further versions of the Observations Data Model were circulated in April, June and October 2006. These documented changes made in the evolution of this design. The fundamental design, however, 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. This document describes the first release version of the data model design, which has been named ODM Release Version 1.0, and has been so named to correspond to the Version 1.0 release of the CUAHSI HIS. This document supersedes the previous documents.

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 database management 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).

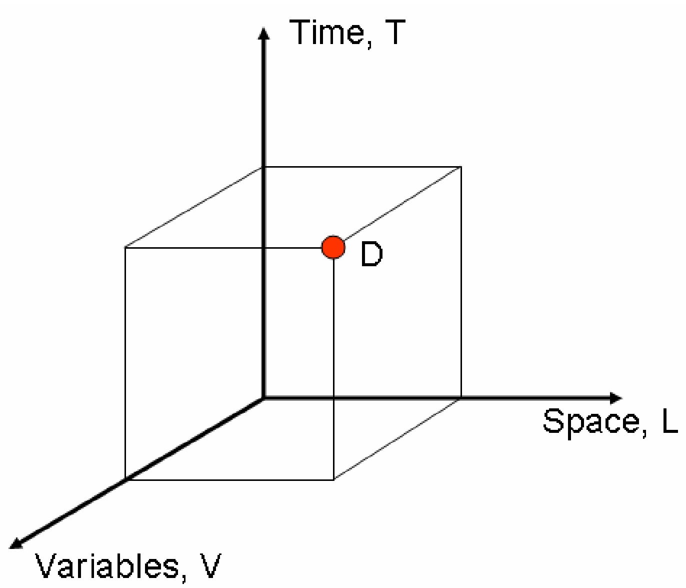


Figure 1. A measured data value (D) is indexed by its spatial location (L), its time of measurement (T), and what kind of variable it is (V).

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, UT"), 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
Data Value	The observation value itself
Accuracy	Quantification of the measurement accuracy associated with the observation value
Date and Time	The date and time of the observation (including time zone offset relative to UTC and daylight savings time factor)
Variable Name	The name of the physical, chemical, or biological quantity that the data value represents (e.g. streamflow, precipitation, temperature)
Location	The location at which the observation was made (e.g. latitude and longitude)
Units	The units (e.g. m or m3/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)
Offset Type/ 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. a continuous, minimum, maximum, 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/ Method	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 organization, agency, 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 data 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 DataValues 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 specify the terms that may be used for the fields where the vocabulary is controlled.

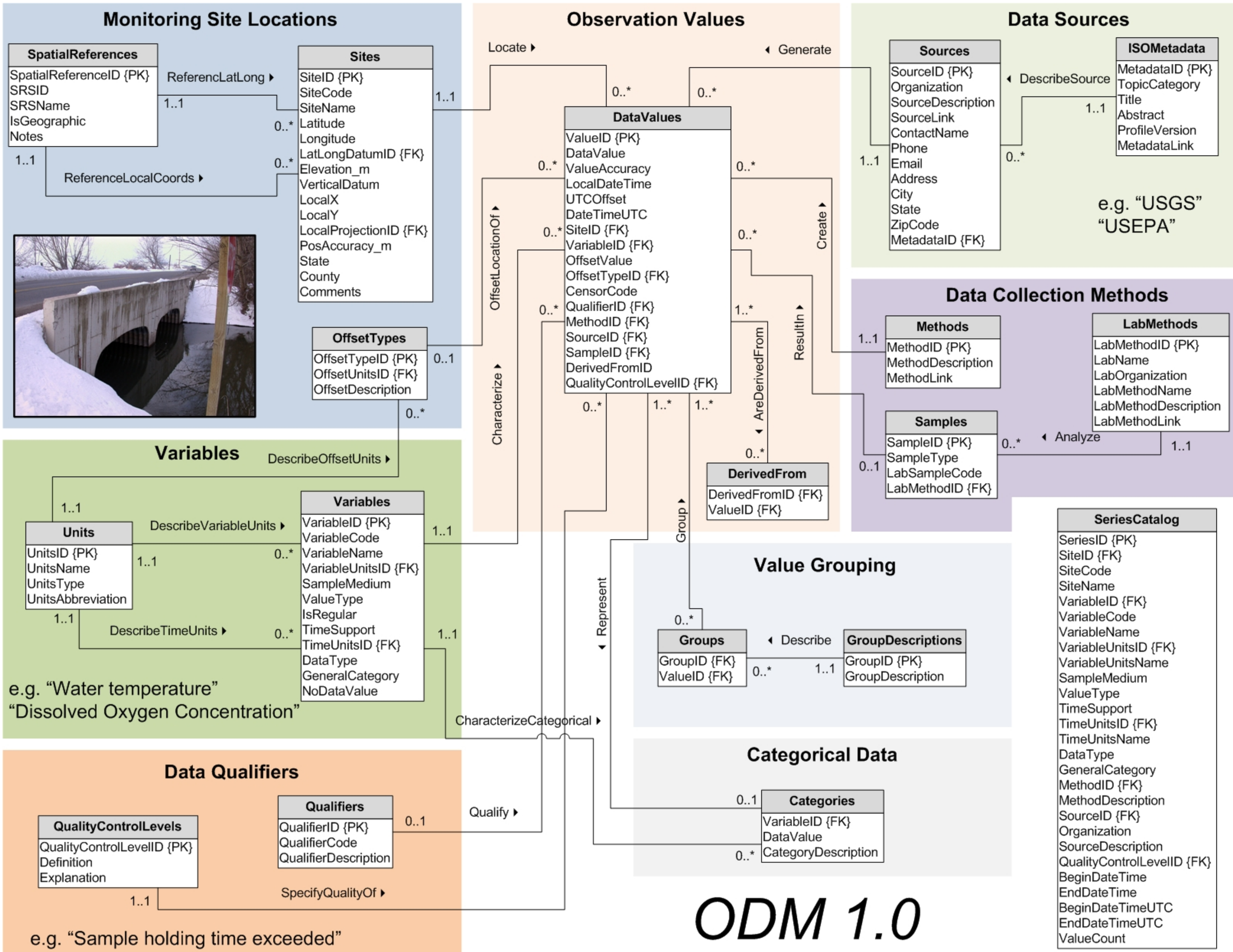


Figure 2. Observations Data Model schema.

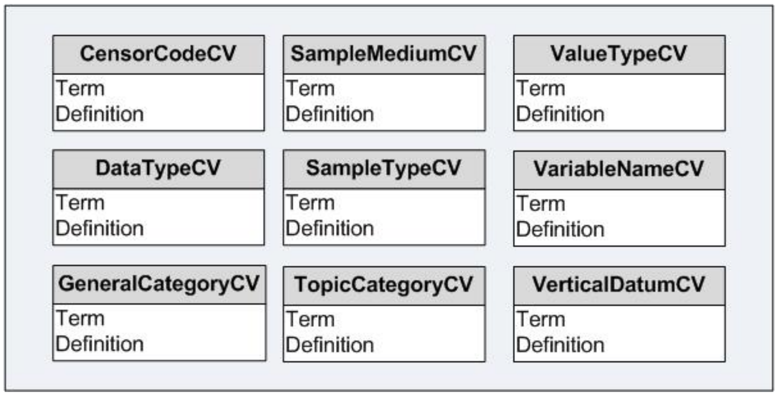


Figure 3. Controlled vocabulary tables.

Table 2. Observations Data Model Logical Relationships

Relationships that define ancillary information about data values

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

Relationships that define derived from groups

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

Relationships that define groups

Table	Field	Type	Field	Table
DataValues	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
DataValues	DataValue	* <-> 1	DataValue	Categories

Relationships used to define the Units

Table	Field	Type	Field	Table
Units	UnitsID	1<->*	VariableUnitsID	Variables
Units	UnitsID	1<->*	TimeUnitsID	Variables
Units	UnitsID	1<->*	OffsetUnitsID	OffsetTypes

Relationship used to define the <u>Sample Laboratory Methods</u>

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

Relationship used to define the <u>Spatial References</u>

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

Relationship used to define the <u>ISOMetaData</u>

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 data 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 data values that have been used to derive other data values. The third set of relationships defines logical groupings of data values. The fourth set of relationships is used to specify the categories associated with categorical variables. The fifth set of relationships is used to define the units. The sixth set of relationships associates laboratory methods with samples. The seventh 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 Observations Data Model Design

<u>Geography</u>

The ODM is intended to be independent of the geographical representation of the site locations. Earlier versions of the data model had a “Shape” attribute 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, which 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 complex implementations, and especially when multiple databases are merged into a single ODM, it may not be possible to preserve the simple one-to-one relationship between SiteID and HydroID with each of these fields holding the same integer identifier values. In these cases, where SiteID and HydroID are not the same, a bridge table would be used to associate the ODM SiteIDs used to identify sites with HydroIDs in the Arc Hydro data model.

SiteID must be unique within an instance of 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 and 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 data 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.

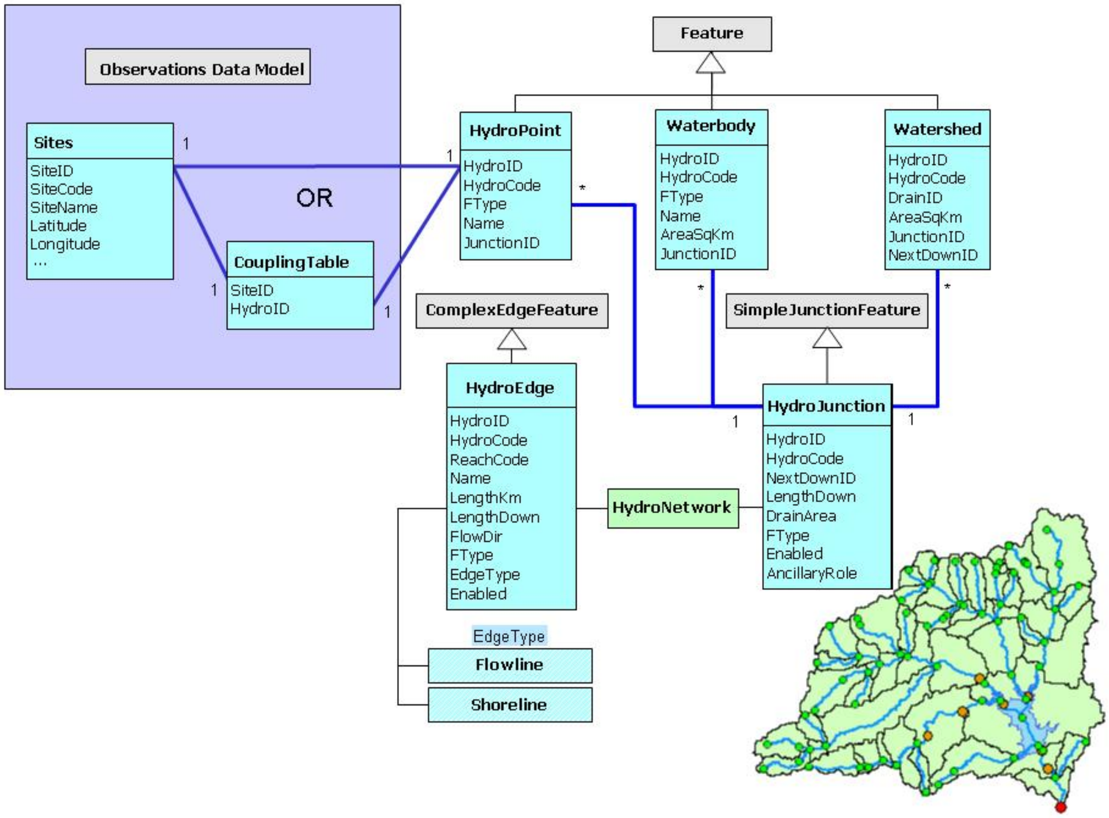


Figure 4. Arc Hydro Framework Data Model and Observations Data Model related through SiteID field in the Sites table.

Series Catalog

A "data series" is an organizing principle that is present in the ODM. A data series consists of all the data values associated with a unique site, variable, method, source, and quality control level combination in the DataValues table. The SeriesCatalog table lists data series identifying each by a unique series identifier, SeriesID. This table is essentially a summary of many of the tables in the ODM and is not required to maintain the integrity of the data. However, it serves to provide a listing of all the distinct series of data values of a specific variable at a specific site. By doing so, this table provides a means by which users can execute most common data discovery queries (i.e., which variables have data at a site, etc.) without the overhead of querying the entire DataValues table, which can become quite large.

The SeriesCatalog table is also intended to support CUAHSI Web Service method queries such as GetSiteInfo, which returns information about a monitoring site within an instance of the ODM including the variables that have been measured at that site. It should be noted that data series, as they are defined here, do not distinguish between different series of the same variable at the same site but measured with different offsets. If for example temperature was measured at two different offsets by two different sensors at one site, both sets of data would fall into one data series for the purposes of the SeriesCatalog table. In these cases, interpretation or analysis software will need to specifically examine and parse the offsets by examining the offset associated with each data value. The SeriesCatalog table does not do this because its principal purpose is data discovery, which we did not want to be overly complicated. The SeriesCatalog table should be programmatically generated and modified as data are added to the database.

Accuracy

Each data value in the DataValues table has an associated attribute called ValueAccuracy. This is a numeric value that quantifies the total measurement accuracy defined as the nearness of a measurement to the true or standard value. Since the true value is not known, the ValueAccuracy is estimated based on knowledge of the instrument accuracy, measurement method, and operational environment. The ValueAccuracy, which is also called the uncertainty of the measurement, compounds the estimates of both bias and precision errors. Bias errors are generally fixed or systematic and cannot be determined statistically, while precision errors are random, being generated by the variability in the measurement system and operational environment. Figure 5 illustrates the effects of these errors on a sample of measurements. Bias errors are usually estimated through specially designed experiments (calibrations). The precision errors are determined using statistical analysis by quantifying the measurement scatter, which is proportional to the standard deviation of the sample of repeated measurements. The total error is obtained by the root-sum-square of the estimates for bias and precision errors involved in the measurement. Figure 6 gives another illustration of the ValueAccuracy concept based on the analogy of a target, where the bulls eye at the center represents the true value.

ValueAccuracy is a data value level attribute because it can change with each measurement, dependent on the instrument or measurement protocol. For example, if streamflow is measured using a V-notch weir, it is actually the stage that is measured, with accuracy limited by the precision and bias of the depth recording instrument. The conversion to discharge through the stage-discharge relationship results in greater absolute error for larger discharges. Inclusion of the ValueAccuracy attribute, which will be blank for many historic datasets because historically accuracy has not been recorded, adds to the size of data in the ODM, but provides a way for factoring the accuracy associated with measurements into data analysis and interpretation, a practice that should be encouraged.

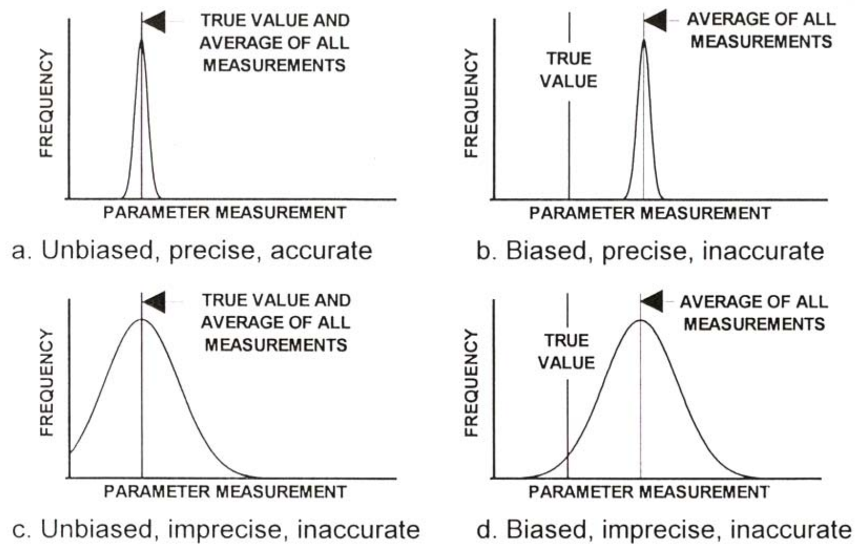


Figure 5. Illustration of measurement error effect (Source: AIAA, 1995).

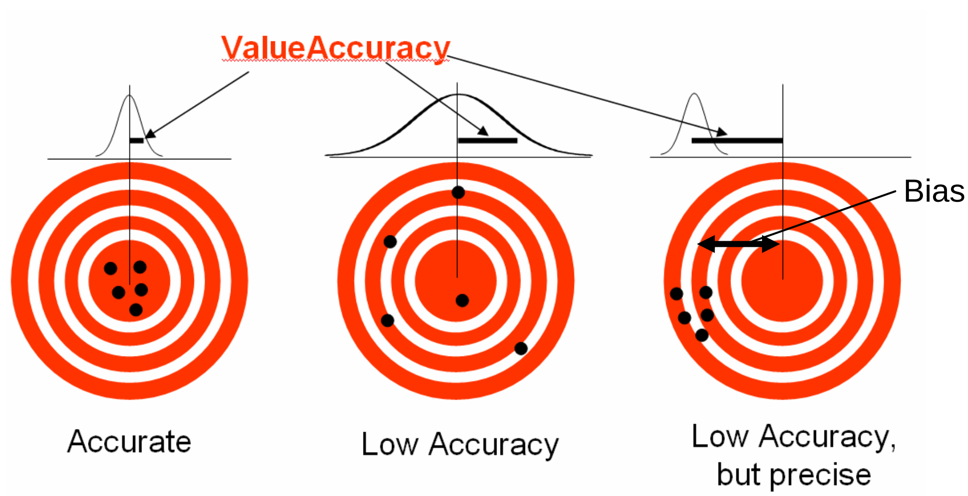


Figure 6. Illustration of Accuracy versus Precision (adapted from Wikipedia http://en.wikipedia.org/wiki/Accuracy).

In designing this ODM, consideration was given to the suggestion by some reviewers to record bias and precision separately, in addition to ValueAccuracy for each data value. This has not been done at this release in the interest of parsimony and also because quantifying these separate components of the error is difficult. We suggest that for most measurements there should be the presumption that they are unbiased and that ValueAccuracy quantifies the precision and accuracy in the judgment of the investigator responsible for collecting the data. For cases where there is specific bias and precision information to complement the ValueAccuracy attribute, this could be recorded in the ODM as a separate variable, e.g. discharge precision, or temperature bias. The groups and derived from features (see below) could be used to associate these variables with their related observations. For measurements that are known to be biased, we suggest that the bias could be quantified by other reference measurements that should also be placed in the database and that a new set of corrected measurements that have had the bias removed should be added to the database at a higher quality control level. These new measurements should have a lower ValueAccuracy value to reflect the improvement in accuracy by removal of the bias. The method and derived from information for these corrected measurements should give the bias removal method and refer to the data used to quantify and remove the bias.

Offset

Each record in the DataValues table has two optional fields OffsetValue 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 an OffsetValue into an OffsetTypes table that gives units and definition associated with the OffsetValue. This design only has the capability to represent one offset for each data value. 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 used for latitude, longitude, and elevation be specified. The SpatialReferences 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 SpatialReferences table Notes field. The accuracy with which the location of a monitoring site is known is quantified using the PosAccuracy_m field in the Sites table. 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 coordinates) in meters. Using a large number for PosAccuracy_m (e.g. 2000 m) accommodates entry of data collected for a study area where the precise location where the observation was recorded is not known.

Groups and Derived from Associations

The DerivedFrom and Groups tables fulfill the function of grouping data 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 an 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 DataValues table (through the DerivedFromID field in that table). This record would have been derived from the data values identified by the group. The method of derivation would be given through the methods table associated with the data value. This construct is useful, for example, to identify the 96 15-minute unit streamflow values that go into the estimate of the mean daily streamflow. Note that there is no limit to how many groups a data value may be associated with, and data values that are derived from other data values may themselves belong to groups used to derive other data values (e.g. the daily minimum flow over a month derived from daily values derived from 15 minute unit values). Note also that a derived from group may have as few as one data value for the case where a data value is derived from a single more primitive data value (e.g., discharge from stage). Through this construct the ODM has the capability to store raw observation values and information derived from raw observations, while preserving the connection of each data 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 the Groups table that have the same GroupID, but different ValueID, that serve to define the group. In Figure 1, the Group relationship is labeled 1..*, at the DataValues table and 0..* at the Groups table. This indicates that a group may comprise one or more data values and that a data value may be included in 0 or more groups. Similarly, there will be many records in the DerivedFrom table that have the same DerivedFromID, but different ValueID that serve to define the group of data values from which a data value is derived. Logically a data 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 data value that has a DerivedFromID should describe the method used for deriving the particular data value from other data values (e.g. calculating discharge from a number of velocity measurements across a stream). The relationship between the DataValues table DerivedFromID field and DerivedFrom table DerivedFromID field is many-to-many (*<->*) because it can occur that the same group of data values is used to derive more than one derived data value. In Figure 1, the AreDerivedFrom relationship between the data values and DerivedFrom table actually depicts both relationships between these tables listed in table 2. The AreDerivedFrom relationship is labeled 1..* at the DataValues table and 0..* at the DerivedFrom table to indicate that a derived from group may comprise 1 or more data values and that a data value may be a member of 0 or more derived from groups.

Dates and Times

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 DataValues 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 data values collected in different time zones (e.g. compare data 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 UTC time, while others may use local time adjusting for daylight saving 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 DateTimeUTC is also included as a record level attribute associated with each data value. This provides a consistent time for querying and sorting data values. There is a level of redundancy between LocalDateTime, UTCOffset and DateTimeUTC. 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, LocalDateTime 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 the date and time 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 as a separate attribute. 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 LocalDateTime and UTCOffset to users so that time may be unambiguously interpreted. In the SeriesCatalog table, begin and end times for each data series are represented by the attributes BeginDateTime, EndDateTime, BeginDateTimeUTC, and EndDateTimeUTC. The UTC offset may be derived from the difference between the UTC and local times. Because local time may change (e.g. with daylight savings) it is important during the derivation of the SeriesCatalog table that identification of the first and last records be based on UTC time and that local times be read from the corresponding records, rather than using a min or a max function on local times which can result in an error.

Support Scale

In interpreting data values that comprise a time series it is important to know the scale information associated with the data values. Blöschl and Sivapalan (1995) review the important issues. Any set of data values is quantified by a scale triplet comprising support, spacing, and extent as illustrated in Figure 7.

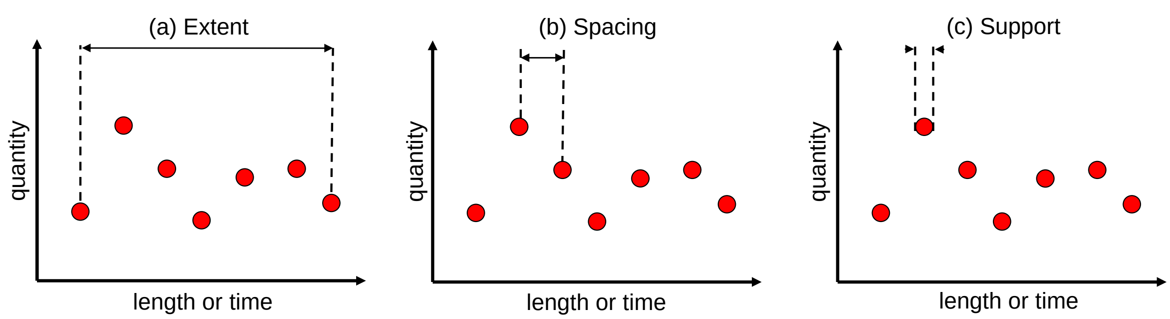


Figure 7. The scale triplet of measurements (a) extent, (b) spacing, (c) support (from Blöschl, 1996).

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 ODM, extent and spacing are properties of multiple measurements and are defined by the LocalDateTime or DateTimeUTC associated with data values. We have included a field called TimeSupport in the Variables table to explicitly quantify support. Figure 8 shows some of the implications associated with support, spacing, and extent in the interpretation of time series data values.

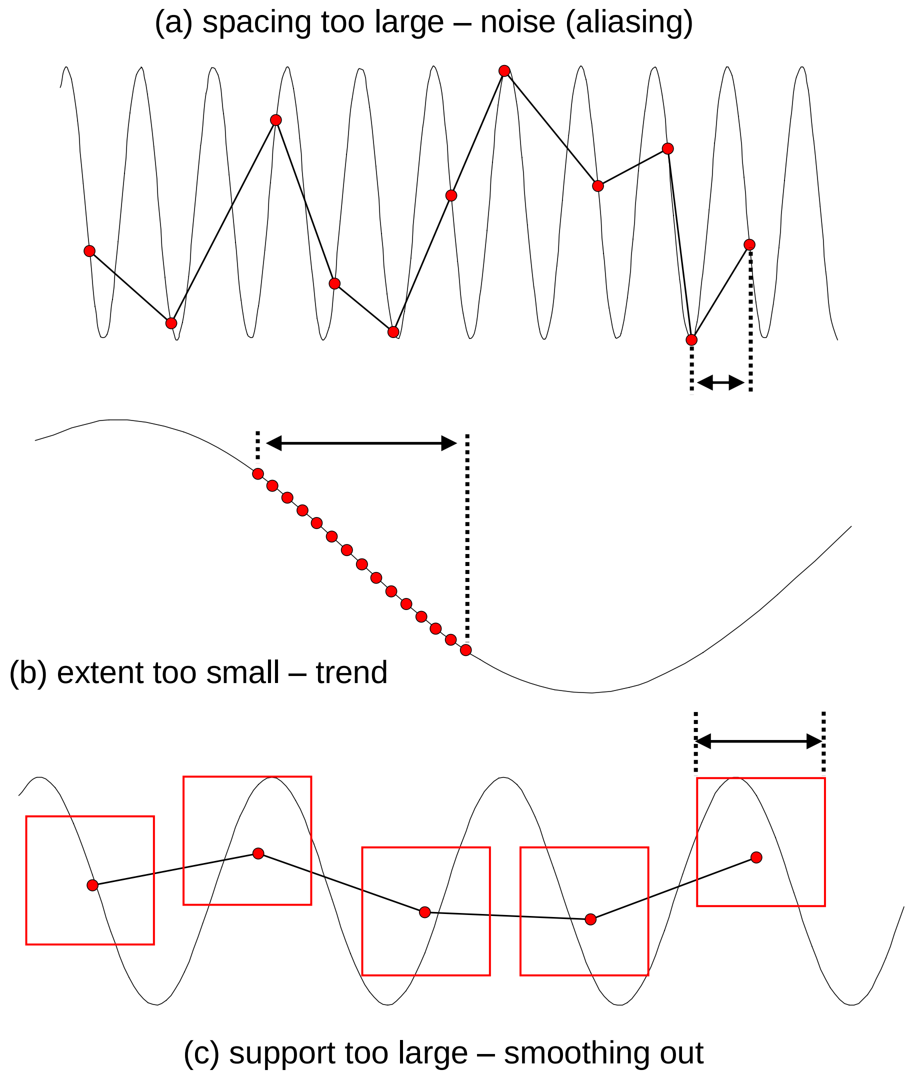


Figure 8. The effect of sampling for measurement scales not commensurate with the process scale: (a) spacing larger than the process scale causes aliasing in the data; (b) extent smaller than the process scale causes a trend in the data; (c) support larger than the process scale causes excessive smoothing in the data (adapted 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. The next section indicates how time support should be specified for the different types of data.

Data Types

In the ODM, 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. Time support may be specified as 0 if the measurements are instantaneous, or given a value that represents the time averaging inherent in the measurement method or device.
2. *Sporadic* 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. As for continuous data, time support may be specified as 0 if the measurements are instantaneous, or given a value that represents the time averaging inherent in the measurement method or device.
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. In the ODM we adopt 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 date and time of the zero record and the current record at the same site position. Site position is defined by a unique combination of SiteID, VariableID, OffsetValue 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. The time support for a cumulative value should be specified as 0 if the measurement of the cumulative quantity is instantaneous, or given a value that represents the time averaging inherent in the measurement of the cumulative value at the end of the period of accumulation.
4. *Incremental* data – the data value represents the incremental value of a variable over a time interval Δt such as the incremental volume of flow, or incremental precipitation: $\Delta V(t) = \int_t^{t+\Delta t} Q(\tau)d\tau$. As for cumulative data, unambiguous interpretation requires knowledge of the time increment. In the ODM we adopt the convention of using TimeSupport if this is given, or the time interval from the previous data value at the same position if TimeSupport is not given or is 0 to specify the interval Δt. This accommodates incremental type precipitation data that is only reported when the data value is non-zero, such as NCDC data. Such NCDC data is irregular, with the interpretation that precipitation is 0 if not reported unless qualifying comments designate otherwise. See example E.4 below for an illustration of how NCDC precipitation data is accommodated in the ODM.
5. *Average* data – the data 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 data value at the same position for irregular data.

6. *Maximum* data – the data 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. The ODM adopts the convention that the time interval is the TimeSupport for regular data and the time interval from the previous data value at the same position for irregular data.
7. *Minimum* data – the data 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 data value is a quantity that can be interpreted as constant over the time interval from the previous measurement.
9. *Categorical* data – the data value is a categorical rather than continuous valued quantity. Mapping from data values to categories is through the Categories table.

We anticipate that additional data types such as median, standard deviation, variance may need to be added as users start to work with ODM.

<u>Beginning of Interval Reporting Time for Interval Data Values</u>

Data types 4 to 8 above apply to data values that occur over an interval of time. The date and time reported and entered in to the ODM database associated with each interval data value is the beginning time of the observation interval. This convention was adopted to be consistent with the way dates and times are represented in most common database management systems. It should be noted that using the beginning of the interval is not consistent with the time a data logger would log an observation value. Care should be exercised in adding data to the ODM to ensure that the beginning of interval convention is followed.

<u>Time Series Data</u>

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,” which is designed to indicate that the data 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 data values as “Continuous.” For example, monthly grab samples of water quality are not continuous, but are better categorized as having DataType “Sporadic.” Note that ODM 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 can be easily computed from date and time values by analysis tools. The inference of measurement gaps (and what to do about them) from date and time values we also regard as analysis functionality left for a Hydrologic Analysis System to handle.

Categorical Variables

In the ODM categorical or ordinal variables are stored in the same table as continuous valued 'real' variables through a numerical encoding of the categorical data value as a 'real' data value. The Categories table then associates, for each variable, a data 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 DataValues table for categorical data values. The Categories table stores the categories associated with categorical data 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 DataValues and Categories. The CategoryDescription 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 data values are derived from a physical sample that is later analyzed in a laboratory (e.g. a water chemistry sample or biological sample). The 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 observation was made or 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, dissolved oxygen probe, humidity sensor, or temperature sensor.

Data Qualifiers

Each record in the DataValues table has an attribute called QualifierID that references the Qualifiers table. Each QualifierID in the Qualifiers table has attributes QualifierCode and QualifierDescription that provide qualifying information that can note anything unusual or problematic about individual observations such as for example, "holding time for analysis exceeded" or "incomplete or inexact daily total." Specification of a QualifierID in the DataValues table is optional, with the inference that if a QualifierID is not specified then the corresponding data value is not qualified.

Quality Control Level Encoding

While data qualifiers are value level attributes, each data value in the DataValues table has an attribute called QualityControlLevelID that is designed to record the level of quality control processing that the data value has been subjected to at the level of data series, and as such is one of the attributes (together with site, variable, method, and source) used to uniquely identify data series. QualityControlLevelID is an integer value between 0 and 4. The following level definitions are adapted from those used by other similar systems, such as NASA, Earthscope and Ameriflux (e.g. http://ilrs.gsfc.nasa.gov/reports/ilrs_reports/9809_attach7a.html, http://public.ornl.gov/ameriflux/available.shtml accessed 3/6/2007) and are suggested so that CUAHSI ODM is consistent with the practice of other data systems:

- **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 stored in the QualityControlLevels table. Appendix B of this document provides a discussion of how to handle data versioning in terms of quality control levels, data series editing, and data series creation within ODM 1.0

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. In its design, the ODM embodies much record, variable, and site level metadata. Dataset and project level metadata required by these standards, such as TopicCategory, Title, and Abstract are included in a table called ISOMetaData linked to each data source.

Reference Documents

The Methods, Sources, LabMethods and ISOMetaData tables contain fields that can be used to store 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 URLs included are stable. An alternative approach to external links is to exploit the capability of modern databases to store entire digital documents, such as an html or xml page, PDF document, or raw data file, within a field in the database. The capability therefore exists to instead have these links refer to a separate table that would actually contain this metadata information, instead of housing it in a separate digital library. There is some merit in this because then any data exported in ODM format could take with it the associated metadata required to completely define it as well as the raw data upon which it is derived. However, this 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:

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

We have specified the initial contents of these controlled vocabularies in the accompanying Microsoft SQL Server 2005 schema for the ODM and spreadsheet file[3]. However, we anticipate 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 model to store different types of point observations. It is not possible in examples such as this to present all of the field values for all the tables. Because of this, the examples present selected fields and tables chosen to illustrate key capabilities of the data model. Refer to Appendix A for the complete definition of table and field contents.

[3] Current versions of these files are maintained in the Hydrologic Information System section of the CUAHSI website http://www.cuahsi.org, together with the most recent version of this document.

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 ODM. Note that gage height in feet and discharge in cubic feet per second are both in the same data table but with different VariableIDs that reference the Variables table, which specifies the Variable, Units, and other quantities associated with these data values. The link between VariableID in the DataValues 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 DataValues table references into the DerivedFrom table that references back to the corresponding gage height in the DataValues table from which the discharge was derived.

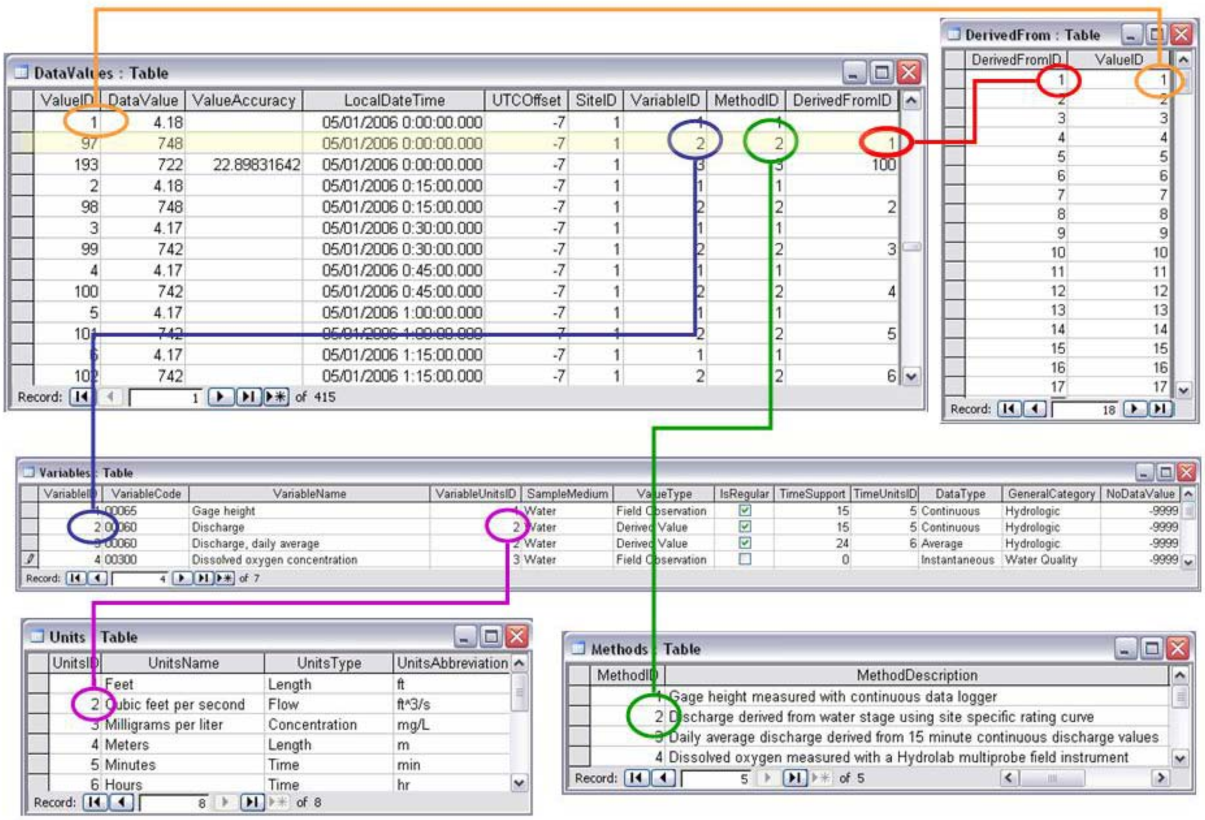


Figure E.1. Excerpts from tables illustrating the population of the ODM with streamflow gage height (stage) and discharge data.

Streamflow - Daily Average Discharge

Daily average streamflow is reported as an average of continuous 15 minute interval data values. Figure E.2 shows excerpts from tables illustrating the population of the ODM with both the continuous discharge values and derived daily averages. The record giving the single daily average discharge with a value of 722 ft^3/s in the DataValues table has a 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 DataValues table used to derive the average daily discharge. VariableID in the DataValues table identifies the appropriate record in the Variables table specifying that this is a daily average discharge with units of ft^3/s from UnitsID referencing in to the Units table. MethodID in the DataValues table identifies the appropriate record in the Methods table specifying that the method used to obtain this data value was daily averaging.

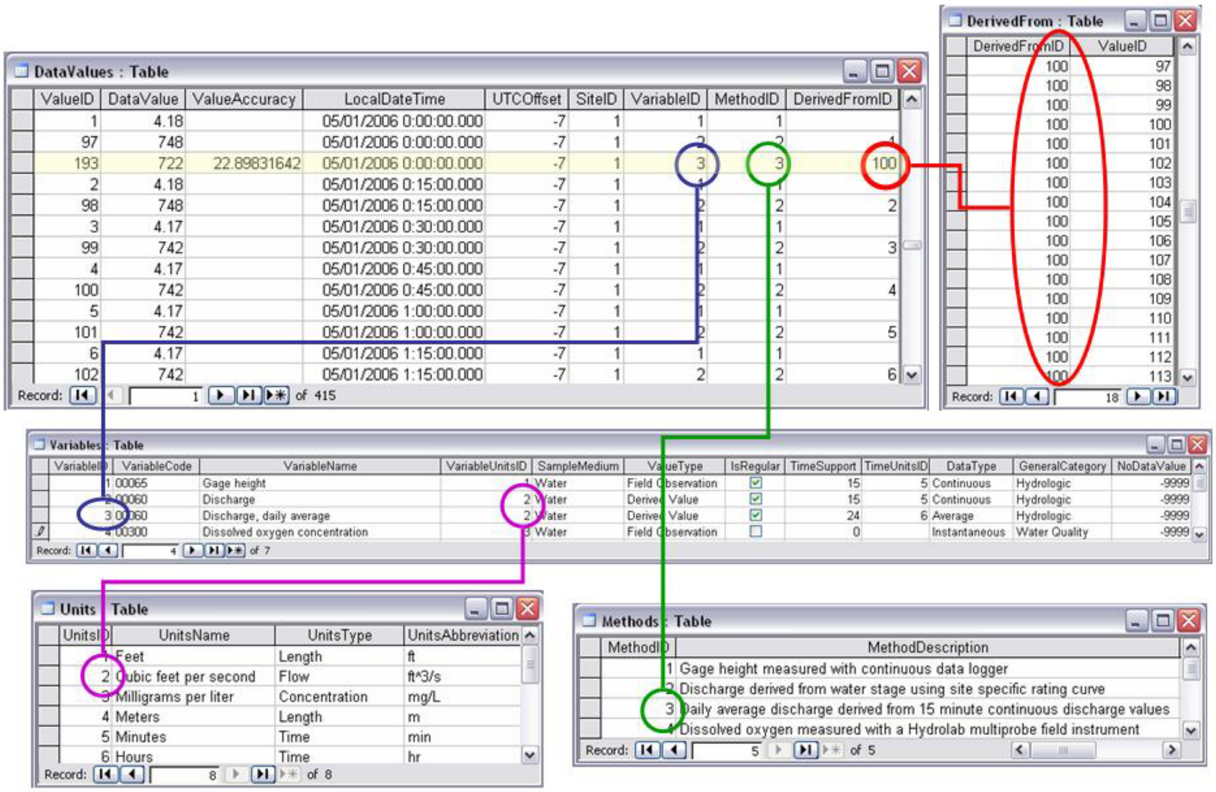


Figure E.2. Excerpts from tables illustrating the population of the ODM with daily average discharge derived from 15 minute discharge values.

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

Reservoir profile measurements provide an example of the logical grouping of data values and data values that have an offset in relationship to the location of the monitoring site. 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 ODM. The OffsetTypes table and OffsetValue attribute is used to quantify the depth offset associated with each measurement. Each of the data values shown has an OffsetTypeID that references into the OffsetTypes table. The OffsetTypes table indicates that for this OffsetType the offset is “Depth below water surface.” The OffsetTypes table references into the Units table indicating that the OffsetUnits are meters, so OffsetValue in the DataValues table is in units of meters depth below the water surface. Each of the data 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 data values shown also has a MethodID that in the Methods table indicates that dissolved oxygen was measured with a Hydrolab multiprobe. The data values shown are part of a logical group of data 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 data values belonging to the group. A description of this group is given in the GroupDescriptions table.

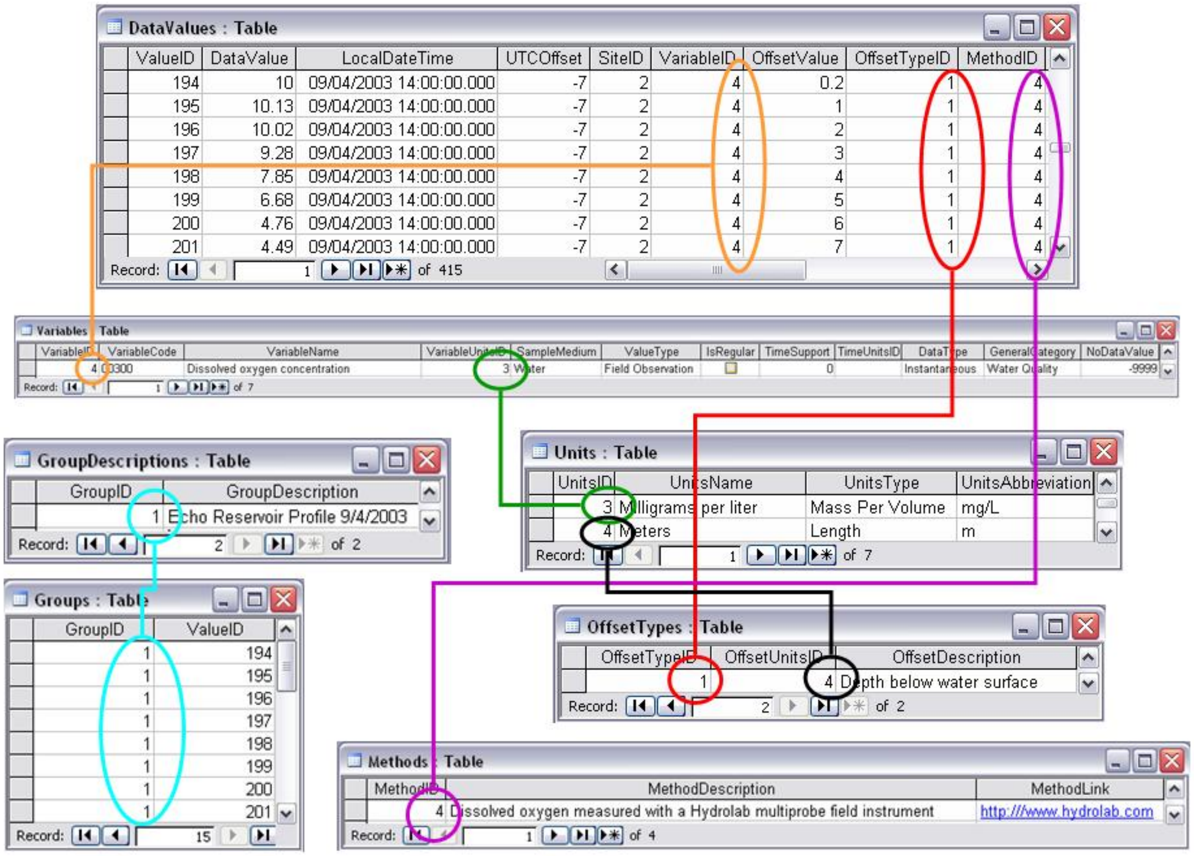


Figure E.3. Excerpts from tables illustrating the population of the ODM with water chemistry data.

<u>NCDC Precipitation Data</u>

Figure E.4 illustrates the representation of NCDC 15 minute precipitation data by the ODM. The data includes 15 minute incremental data values as well as daily totals. Separate records in the Variables table are used for the 15 minute or daily total values. These data are 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 TimeUnits as “Minutes” or “Hours”. The DataType of “Incremental” is used to indicate that these are incremental data 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 data value at the same site position. Data qualifiers 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 is also used to flag days where the precipitation total is incomplete due to the record being missing during part of the day.

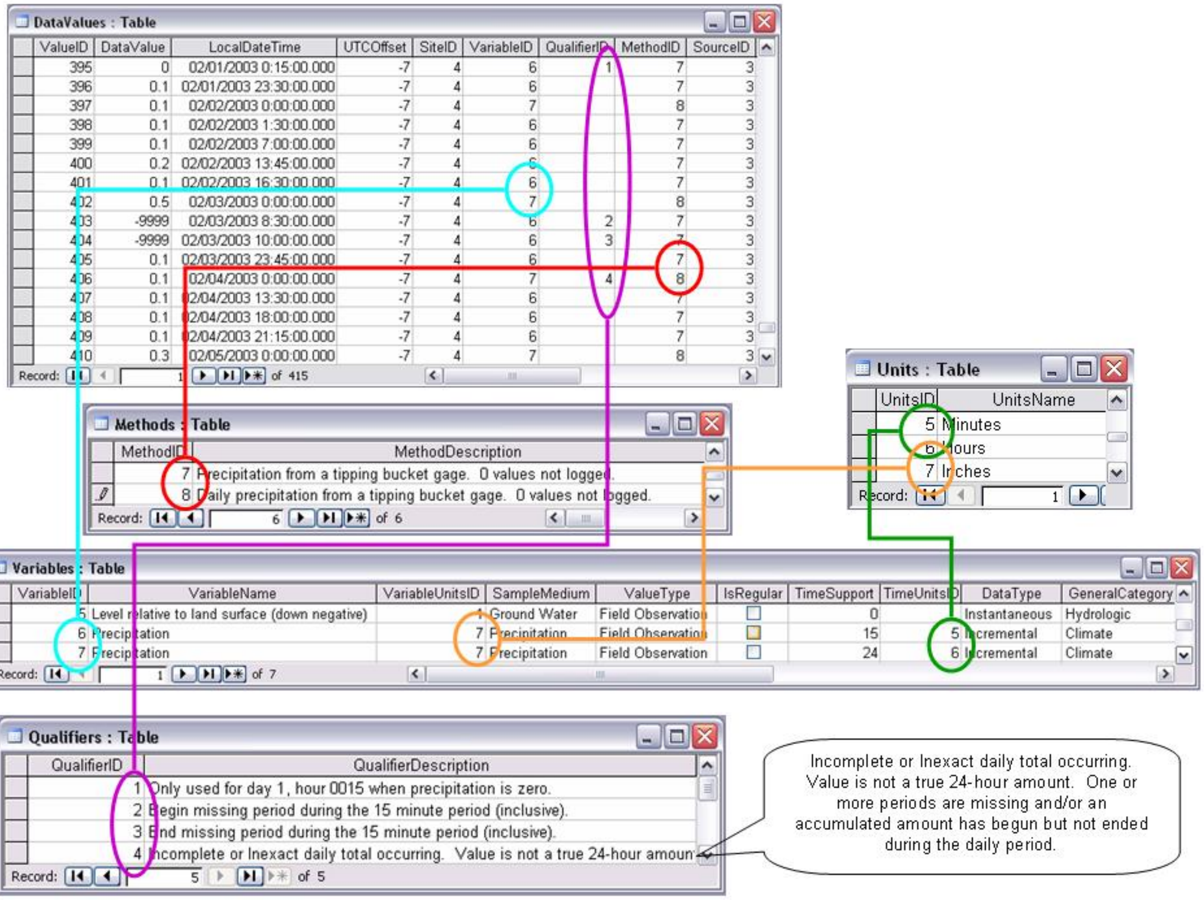
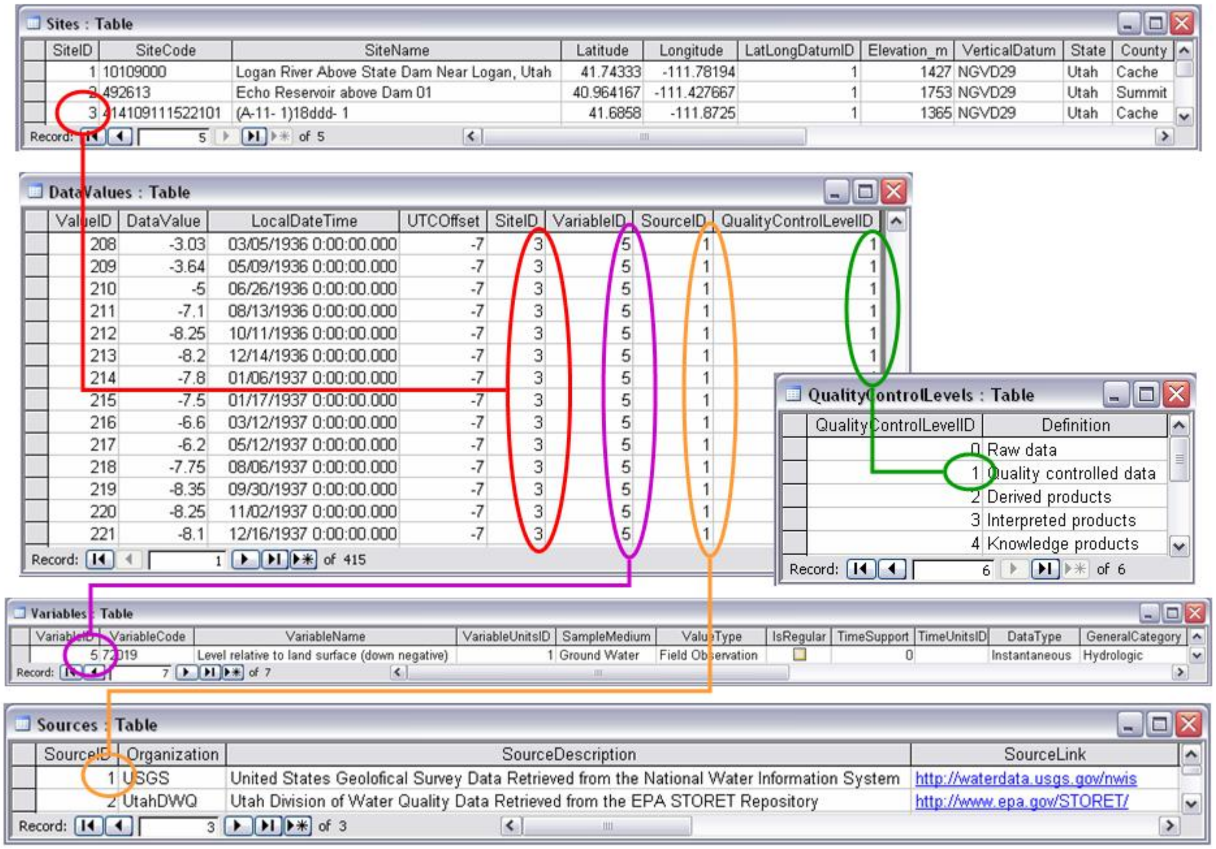


Figure E.4. Excerpts from tables illustrating the population of the ODM with NCDC Precipitation Data.

Groundwater Level Data

The following is an example of how groundwater level data can be stored in the ODM. In this groundwater level example, the data values are the water table level relative to the ground surface reported as negative values. This example shows multiple data values of a single variable at a single site made by a single source that have been quality controlled as indicated by the QualityControlLevelID field in the QualityControlLevels table. The SiteID field in the DataValues table indicates the site in the Sites table that gives the location information about the monitoring 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 DataValues table references the appropriate record in the Variables table indicating information about the variable. The SourceID field in the DataValues table references the appropriate record in the Sources table giving information about the source of the data.



Sites : Table

SiteID	SiteCode	SiteName	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

Record: 5 of 5

DataValues : Table

ValueID	DataValue	LocalDateTime	UTCOffset	SiteID	VariableID	SourceID	QualityControlLevelID
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	

Record: 1 of 415

QualityControlLevels : Table

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

Record: 6 of 6

Variables : Table

VariableID	VariableCode	VariableName	VariableUnitsID	SampleMedium	ValueType	IsRegular	TimeSupport	TimeUnitsID	DataType	GeneralCategory
5	72019	Level relative to land surface (down negative)	1	Ground Water	Field Observation		0		Instantaneous	Hydrologic

Record: 7 of 7

Sources : Table

SourceID	Organization	SourceDescription	SourceLink
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/

Record: 3 of 3

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

References

AIAA, (1995), Assessment of Wind Tunnel Data Uncertainty, American Institute of Aeronautics and Astronautics: AIAA S-071-1995.

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/his/docs/HISStatusSept15.pdf.

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

Maidment, D. R., (2005), "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.

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

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. 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

Table that defines the categories for categorical variables. Records are required for variables where DataType is "Categorical". Multiple entries for each VariableID, with different DataValues provide the mapping from DataValue 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
DataValue	Real	Numeric value	1.0	M
CategoryDescription	Text	Definition of categorical variable value	“Cloudy”	M

Table: CensorCodeCV

Table that contains the 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

Field Name	Data Type	Description	Examples	Validation
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: DataValues

Table that contains the actual data values.

Field Name	Data Type	Description	Example	Validation
ValueID	Integer	Unique integer identifier for each data value	43	M
DataValue	Real	The numeric value of the observation. For Categorical variables, a number is stored here. The Variables table has DataType as Categorical and the Categories table maps from the DataValue onto Category Description.	34.5	M
ValueAccuracy	Real	Numeric value that describes the measurement accuracy of the data value. If not given, it is interpreted as unknown.	4	O
LocalDateTime	Date/Time	Local date and time at which the data 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 of the corresponding LocalDateTime value.	-7	M
DateTimeUTC	Date/Time	Universal UTC date and time at which the data value was observed. Represented in an implementation specific format.	9/4/2003 2:00:00 PM	M
SiteID	Integer	Unique integer identifier for each sampling location. This links data values to their locations in the Sites table.	3	M
VariableID	Integer	Integer identifier that references the variable that was measured. This links data values to their type in the Variables table.	5	M
OffsetValue	Real	Distance from a datum or control point to the point at which a data value was observed. If not given the OffsetValue is inferred to be 0, or not relevant/necessary.	2.1	O
OffsetTypeID	Integer	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.	"nc"	M

Field Name	Data Type	Description	Example	Validation
QualifierID	Integer	Integer identifier that references the Qualifiers table. If Null, the data value is inferred to not be qualified.	4	O
MethodID	Integer	Integer identifier that references method used to generate the data value in the Methods table.	3	M
SourceID	Integer	Integer identifier that references the record in the Sources table giving the source of the data value.	5	M
SampleID	Integer	Integer identifier that references into the Samples table. This is required only if the data value resulted from a physical sample processed in a lab.	7	O
DerivedFromID	Integer	Integer identifier for the derived from group of data values that the current data value is derived from. This refers to a group of derived from records in the DerivedFrom table. If null or not given. the data value is inferred to not be derived from another data value.	5	O
QualityControlLevelID	Integer	Integer between 0 and 4 inclusively giving the level of quality control that the value has been subjected to. This references the QualityControlLevels table.	1	M

Table: DerivedFrom

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

Field Name	Data Type	Description	Examples	Validation
DerivedFromID	Integer	Integer identifying the group of data values from which a quantity is derived.	3	M
ValueID	Integer	Integer identifier referencing data values that comprise a group from which a quantity is derived. This corresponds to ValueID in the DataValues 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

Table that lists the descriptions for each of the groups that have been formed.

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

Table: Groups

Table that lists the groups of data values that have been created and the data values that are within each group.

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

Table: ISOMetadata

Table that contains dataset and project level metadata required by the CUAHSI HIS metadata system (http://www.cuahsi.org/his/documentation.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
MetadataLink	Hyperlink	Link to additional metadata reference material.		O

Table: LabMethods

Table that contains descriptions of the laboratory methods used to analyze physical samples for specific constituents.

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
LabOrganization	Text	Organization responsible for sample analysis.	"USGS"	M
LabMethodName	Text	Name of the method and protocols used for sample analysis.	"USEPA-365.1"	M
LabMethodDescription	Text	Description of the method and protocols used for sample analysis.	"Processed through Model *** Mass Spectrometer"	M
LabMethodLink	Hyperlink	Link to additional reference material on the analysis method.		O

Table: Methods

Table that lists the methods used to collect the data and any additional information about the method such as 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
MethodDescription	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
MethodLink	Hyperlink	Link to additional reference material on the method.		O

Table: OffsetTypes

Table that 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
OffsetUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the OffsetValue.	1	M

Field Name	Data Type	Description	Example	Validation
OffsetDescription	Text	Full text description of the offset type.	“Below water surface” “Above Ground Level”	M

Table: Qualifiers

Table that contains data qualifying comments that accompany the data.

Field Name	Data Type	Description	Example	Validation
QualifierID	Integer	Unique Integer identifying the data qualifier.	3	M
QualifierCode	Text	Text code used by organization that collects the data.	“e” (for estimated) or “a” (for approved) or “p” (for provisional)	O
QualifierDescription	Text	Text of the data qualifying comment.	“Holding time for sample analysis exceeded”	M

Table: QualityControlLevels

Table that contains the controlled vocabulary for quality control levels.

Field Name	Data Type	Description	Example	Validation
QualityControlLevelID	Integer	Integer between 0 and 4 inclusively giving the level of quality control that the data 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.”	M

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

Table that gives information about physical samples analyzed in a laboratory.

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”, “Grab Sample”	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 the sample. This references 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”, “Grab Sample”	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

Table that lists each separate data series in the database for the purposes of identifying or displaying what data are available at each site and to speed simple queries without querying the main DataValues table. Unique site/variable combinations are defined by unique combinations of SiteID, VariableID, MethodID, SourceID, and QualityControlLevelID.

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
SiteID	Integer	Site identifier from the Sites table.	7	P
SiteCode	Text	Site code used by organization that collects the data.	“1002000”	P
SiteName	Text	Full text name of sampling site.	“Logan River”	P
VariableID	Integer	Integer identifier for each Variable that references the Variables table.	4	P
VariableCode	Text	Variable code used by the organization that collects the data.	“00060”	P
VariableName	Text	Name of the variable from the variables table.	“Temperature”	P

Field Name	Data Type	Description	Example	Validation
VariableUnitsID	Integer	Integer identifier that references the record in the Units table giving the Units of the data value.	5	P
VariableUnitsName	Text	Full text name of the variable units from the UnitsName field in the Units table.	“milligrams per liter”	P
SampleMedium	Text	The medium of the sample. This should be from the SampleMediumCV controlled vocabulary table.	“Surface Water”	P
ValueType	Text	Text value indicating what type of data value is being recorded. This should be from the ValueTypeCV controlled vocabulary table.	“Field Observation”	P
TimeSupport	Real	Numerical value that indicates the time support (or temporal footprint) of the data values. 0 is used to indicate data values that are instantaneous. Other values indicate the time over which the data values are implicitly or explicitly averaged or aggregated.	0, 24	P
TimeUnitsID	Integer	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	P
TimeUnitsName	Text	Full text name of the time support units from the UnitsName field in the Units table.	“hours”	P
DataType	Text	Text value that identifies the data as one of several types from the DataTypeCV controlled vocabulary table.	“Continuous” “Instantaneous” “Cumulative” “Incremental” “Average” “Minimum” “Maximum” “Constant Over Interval” “Categorical”	P
GeneralCategory	Text	General category of the variable from the GeneralCategoryCV table.	“Water Quality”	P
MethodID	Integer	Integer identifier that identifies the method used to generate the data values and references the Methods table.	2	P

Field Name	Data Type	Description	Example	Validation
MethodDescription	Text	Full text description of the method used to generate the data values.	"Specific conductance measured using a Hydrolab" or "Streamflow measured using a V notch weir with dimensions xxx"	P
SourceID	Integer	Integer identifier that identifies the source of the data values and references the Sources table.	5	P
Organization	Text	Text description of the source organization from the Sources table.	"USGS"	P
SourceDescription	Text	Text description of the data source from the Sources table.	"Text file retrieved from the EPA STORET system indicating data originally from Utah Division of Water Quality"	P
QualityControlLevelID	Integer	Integer identifier that indicates the level of quality control that the data values have been subjected to.	0,1,2,3,4	P
BeginDateTime	Date/Time	Date of the first data 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 data value in the series. To be programmatically updated if new records are added.	9/4/2005 7:00:00 AM	P
BeginDateTimeUTC	Date/Time	Date of the first data value in the series in UTC. To be programmatically updated if new records are added.	9/4/2003 2:00 PM	P
EndDateTimeUTC	Date/Time	Date of the last data value in the series in UTC. To be programmatically updated if new records are added.	9/4/2003 2:00 PM	P
ValueCount	Integer	The number of data values in the series identified by the combination of the SiteID, VariableID, MethodID, SourceID and QualityControlLevelID fields.	50	P

Table: Sites

Table that provides information giving the spatial location at which data values have been collected.

Field Name	Data Type	Description	Example	Validation
SiteID	Integer	Unique identifier for each sampling location.	37	M
SiteCode	Text	Code used by organization that collects the data to identify the site	“10109000” (USGS Gage number)	O
SiteName	Text	Full name of the sampling site.	“LOGAN RIVER ABOVE STATE DAM, NEAR LOGAN,UT”	M
Latitude	Real	Latitude in decimal degrees.	45.32	M
Longitude	Real	Longitude in decimal degrees. East positive, West negative.	-100.47	M
LatLongDatumID	Integer	Identifier that references the Spatial Reference System of the latitude and longitude coordinates in the 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
VerticalDatum	Text	Vertical datum of the elevation. Controlled Vocabulary from VerticalDatumCV.	“NAVD88”	M
LocalX	Real	Local Projection X coordinate.	456700	O
LocalY	Real	Local Projection Y Coordinate.	232000	O
LocalProjectionID	Integer	Identifier that references the Spatial Reference System of the local coordinates in the SpatialReferences table. This field is required if local coordinates are given.	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 monitoring site is located.	“Utah”	O
County	Text	Name of County in which the monitoring site is located.	“Cache”	O
Comments	Text	Comments related to the site,		O

Table: Sources

Table that lists the original sources of the data, providing information sufficient to retrieve and reconstruct the data 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

Field Name	Data Type	Description	Example	Validation
Organization	Text	Name of the 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
SourceDescription	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
SourceLink	Hyperlink	Link that can be pointed at the original data file and/or associated metadata stored in the digital library or URL of data source.		O
ContactName	Text	Name of the contact person for the data source.	"Jane Adams"	M
Phone	Text	Phone number for the contact person.	"435-797-0000"	M
Email	Text	Email address for the contact person.	"Jane.Adams@dwq.ut"	M
Address	Text	Street address for the contact person.	"45 Main Street"	M
City	Text	City in which the contact person is located.	"Salt Lake City"	M
State	Text	State in which the contact person is located. Use two letter abbreviations for US. For other countries give the full country name.	"UT"	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

Table that provides information about the Spatial Reference Systems used for latitude and longitude as well as local coordinate systems in the Sites table.

Field Name	Data Type	Description	Example	Validation
SpatialReferenceID	Integer	Unique integer identifier for each Spatial Reference System.	37	M
SRSID	Integer	Integer identifier for the Spatial Reference System from http://www.epsg.org/.	4269	O
SRSName	Text	Name of the Spatial Reference System.	"NAD83"	M
IsGeographic	Boolean	Value that indicates whether the spatial reference system uses geographic coordinates (i.e. latitude and longitude) or not.	"True", "False"	M

Field Name	Data Type	Description	Example	Validation
Notes	Text	Descriptive information about the Spatial Reference System. This field would be used to define a non-standard study area specific system if necessary and would contain a description of the 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

Table that gives the Units and UnitsType associated with variables, time support, and offsets. This is a controlled vocabulary table.

Field Name	Data Type	Description	Example	Validation
UnitsID	Integer	Unique integer identifier that identifies each unit.	6	M
UnitsName	Text	Full text name of the units.	"Milligrams Per Liter"	M
UnitsType	Text	Text value that specifies the dimensions of the units.	"Length" "Time" "Mass"	M
UnitsAbbreviation	Text	Text abbreviation for the units.	"mg/L"	M

Table: ValueTypeCV

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

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

Table: VariableNameCV

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

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

Table: Variables

Table that 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
VariableCode	Text	Code used by the organization that collects the data to identify the variable.	“00060” used by USGS for discharge	O
VariableName	Text	Full text name of the variable that was measured, observed, modeled, etc. This should be from the VariableNameCV controlled vocabulary table.	“Discharge”	M
VariableUnitsID	Integer	Integer identifier that references the record in the Units table giving the units of the data values associated with the variable.	4	M
SampleMedium	Text	The medium in which the sample or observation was taken or made. This should be from the SampleMediumCV controlled vocabulary table.	“Surface Water” “Sediment” “Fish Tissue”	M
ValueType	Text	Text value indicating what type of data 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 data values are from a regularly sampled time series.	“True” “False”	M
TimeSupport	Real	Numerical value that indicates the time support (or temporal footprint) of the data values. 0 is used to indicate data values that are instantaneous. Other values indicate the time over which the data values are implicitly or explicitly averaged or aggregated.	0, 24	M

Field Name	Data Type	Description	Example	Validation
TimeUnitsID	Integer	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 values as one of several types from the DataTypeCV controlled vocabulary table.	"Continuous" "Sporadic" "Cumulative" "Incremental" "Average" "Minimum" "Maximum" "Constant Over Interval" "Categorical"	M
GeneralCategory	Text	General category of the data 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 the controlled vocabulary for the VerticalDatum field in the Sites table.

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

Appendix B. Data Versioning Within ODM 1.0

The main text of this document focuses on how ODM is structured to store observations data. It does not address how to manage editing data stored within ODM. Software applications based on ODM will have functionality that will allow data managers and database administrators to modify, delete, change, or otherwise make edits to data stored within ODM. In addition, these software tools will provide functionality to create derived datasets, or datasets that are calculated or derived from data already stored in ODM (i.e., calculate a time series of discharge from a time series of stage, or calculate a time series of daily average temperature from a time series of hourly observations). The purpose of this appendix is to clarify how data editing and versioning can be managed within the ODM 1.0 schema.

Data Series Defined

In order to fully grasp the concepts that follow, the idea of a "data series" in the context of ODM must be clarified. A "data series" is an organizing principle that is present in the ODM. A data series consists of all of the data values associated with a unique site, variable, method, source, and quality control level combination. An example of the full specification for a data series is: "all of the raw unchecked (QualityControlLevel) water temperature (Variable) values measured in the Logan River near Logan, UT (Site) using a field temperature sensor (method) by Utah State University (Source)." Each record in the SeriesCatalog table of ODM represents a unique data series.

Rules for Editing and Deriving Data Series in ODM 1.0

The following rules are suggested so that versioning of and edits to data series can be managed within the ODM 1.0 schema. Software applications that work with ODM should follow these rules.

1. *Data versioning should be done at the data series level* – Within ODM, the concept of data versioning is related to the quality control level. Quality control level is a data series level attribute, and as such, changes to the quality control level should occur at the data series level rather than at the individual value level. For example, if an investigator wished to create a quality controlled Level 1 data series from a raw Level 0 data series, he/she should first make a copy of the raw Level 0 data series and then perform any edits and adjustments required in the quality control process to the copy. The edited copy then becomes the Level 1 data series, and the Level 0 data series is preserved intact.
2. *Data series with a QualityControlLevelID of 0 cannot be edited* – Level 0 data series represent raw data from sensors (i.e., stage measured by a water level recorder) or other products derived from raw data (i.e., discharge that is programmatically derived from stage before the stage values have been quality controlled). By definition, Level 0 data have not been quality controlled and may contain significant errors and bad values. However, Level 0 data series represent the source from which all other derived data series are based, and as such should remain intact for archive purposes. Level 0 data series should not be used for analysis unless no other adequate options are available, and

only if the user is aware that the data are raw. Level 0 data series can be removed entirely from the database, but only by removing the entire data series.

3. *Only one QualityControlLevel 0 data series can exist for a Site, Variable, and Method combination* – Only one raw data series for a Site, Variable, and Method combination can exist within an ODM database. If multiple sensors are measuring the same variable at the same site, the method description would have to distinguish between the two.
4. *Only one QualityControlLevel 1 data series can exist for each Site, Variable, and Method combination* – Once a Level 0 data series has been loaded to the database, a Level 1 data series can be "derived" from that Level 0 data series. This is done by first making a copy of the Level 0 data series, second changing the QualityControlLevel of the copy to 1, and last doing any necessary filtering or editing required so that the Level 1 data series is acceptable as quality controlled. In most cases, the majority of the values within a Level 0 data series and its corresponding Level 1 data series will remain the same. However, where instruments malfunction or other conditions are present that affect the raw data values, Level 0 values may be deleted, adjusted, or otherwise edited in creating the Level 1 data series.
5. *Any edits to a data series are saved to that data series* – Level 0 data cannot be edited. With Levels 1 or higher, however, software applications will be allowed to edit and delete values. Each time an edit is made, the result overwrites the previous value within a data series. In other words, edits do not create new data series, they modify an existing one. This will be true even where edits are done within multiple editing sessions. The editing software should record the method or basis for any data edits in appropriate method records.
6. *Data series of Level 2 or higher can only be created from data series of Level 1 or higher* – Derived data series of Level 2 or higher can only be created from data series of Level 1 or higher. If a user wishes to create a derived data series from a Level 0 data series (such as discharge from raw, unchecked stage values) that derived data series would also be Level 0.