A Data Model for Environmental Observations

Bora Beran[1], David Valentine[2], Catharine van Ingen[1], Ilya Zaslavsky[2], Tom Whitenack[2]
[1]Microsoft Research, [2] San Diego Supercomputer Center

Microsoft Technical Report **MSR-TR-2008-92**
July 3, 2008

Abstract

Hydrologic and other environmental scientists are beginning to use commercial database technologies to locate, assemble, analyze, and archive data. A data model that is both capable of handling the diversity of data and simple enough to be used by non-database professionals remains an open question.

Over the past three years, we have been working in parallel on data models with hydrologists of Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) and Berkeley Water Center (BWC). This report proposes a new data model with learning from both efforts. This new model has major improvements in spatial support, namespace translation, provenance tracking and versioning, extensibility, and supports a wide array of data types.

We have used this data model to build a hydrologic data catalog server (http://www.sciscope.org) containing metadata for over 358 million sensor and point sample measurements.

1. Introduction

The combination of inexpensive sensors, satellites, and internet data availability is creating a flood of data. Moreover, the nature of the data available from these sources is widely variable in time scale, length scale, name convention, and provenance. Data or location properties such as landcover may also be non-numerical or categorical. Environmental data is also inherently spatial.

There are a number of virtual organizations working to help scientists contend with this flood. CASA/LEAD[1], SenseWeb[2] and CUAHSI [3] are three such examples.

CUAHSI is an organization of hydrologists with 115 member institutions. One of the products of CUAHSI's Hydrologic Information System (HIS) project is the Observations Data Model (ODM); this includes a database schema that implements the ODM model used to store hydrologic data. As a part of the NSF Environmental Observatory program [4], 11 testbed sites across the US operated by the WATERS Network [5] and an unknown number of individual researchers implement the products of the HIS project.

Berkeley Water Center [6] is comprised of over 70 faculty members and other researchers from several University of California, Berkeley colleges and departments. With the Sonoma County Water agency, BWC began building a digital watershed of the Russian River to investigate why salmon have become endangered. That work is now expanded to cover 26 additional watersheds in California by a partnership with National Marine Fisheries Service. The Environmental Data Server built by BWC uses a schema similar to the core of ODM.

The CUAHSI approach was top down design; the BWC approach was bottom up just in time implementation. Both approaches are similar to data models used by other efforts. Merging our efforts, we have evolved ODM.

Using both data models, we have come to realize key shortcomings in the initial ODM design.

- Subsetting ODM for specific tasks was difficult.

ODM required that the data be in the database. As such, building subsets for specific purposes was problematic. We discovered that scientists and government agencies are often hesitant to allow a third party to hold and redistribute their data. To simplify data discovery, we needed a metadata catalog describing the available data from distributed data sources but without the actual data. The two large scale implementations of ODM [7],[8] tried to address this issue with somewhat ad hoc modifications.

- The ODM metadata design was cumbersome and often ignored in practice.

We also found that scientists and the computational scientists that support them were using only subsets often out of expediency. This was particularly true for the metadata parts of the data model. Water quality measurements are often the result of a bottle sample; one sample can be subject to a number of different laboratory processes and yield a number of measurements. The metadata necessary to describe that is much larger than for the collection of stream stage measurements from a single gage. ODM lacked clear rules as to what must be included in the metadata.

Our experience indicates that a data model for multidisciplinary science communities should come up with a solution to handle the extensions within the schema rather than trying to answer the question ”How much metadata is enough?”.

- Extending ODM was ad hoc.

We also found that the domain scientists and the computational scientists supporting them were adding tables or columns to the database. Since this was an ad hoc process, there was little or no way that the information could be shared with other ODM users. We had lost the data interchange.

- Vector and raster data were not included. Non-numerical data was poorly defined.

The ODM design center was scalar data originated from point sources such as stationary in-situ sensors or samples. In other words, ODM focused on time series data such as stream gage readings. One of the test bed projects was FERRYMON. The water quality data are obtained on a moving commercial ferry. Remove sensing data such as NEXRAD or MODIS were also not included, yet are of increasing importance to environmental scientists. There were several data types that needed to be addressed.

- Spatial features and spatial data analyses were not well integrated.

Environmental data is often concerned with spatial features such as watershed boundaries, or distance down a river. The initial design relied on extensions by proprietary GIS software which limited the general ability to perform spatial aggregations and other simple calculations on the data.

- ODM lacked pragmatic versioning and name space translation.

Considering the multidisciplinary nature of the environmental sciences, it is almost impossible to come up with a single variable name space. Scientists commonly use different names to refer to the same variable; stream “flow” and “discharge” are examples of this. Some agencies overload the variable name with provenance information including the instrumentation used in the measurement.

During analysis, data may be cleaned, gap-filled, or otherwise corrected. New variables are also derived from combinations of measurements. For example, bed load sediment transport rate may be computed from discharge and channel cross-sections. While ODM included the notion of an individual data point derived from another data point, the more general versioning necessary to support algorithmic data processing was missing. Similarly, there was no ability to tag data as used in a specific publication.

We designed our new data model to address these shortcomings. We implemented the data model using SQL Server 2008; this gives us native support for geographic data and operations.

2. Profiles

The new data model is presented as a series of profiles (Figure 1). By utilizing profiles, core model is intended to be smaller, more manageable, and simpler to understand and populate. Core profile is built around a catalog of observations. This catalog provides a starting point for queries for locating measurements within a spatio-temporal frame. Each entry in the catalog is represented with the geographical location, measured variable, measurement method and the time span. Additional metadata such as publisher, units and measurement frequency are also stored in the catalog. Geoposition attribute utilizes SQL Server’s *geography* datatype which allows representation of features such as watersheds, geologic formations, isohyets, streams or dams in the database which are essentially polygons, polylines or points on a map.

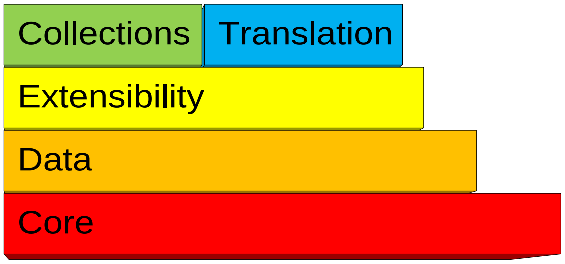


Figure 1 Layers of the Data Model

2.1. Core Layer

Core layer serves the purpose of a central metadata repository whose focus is on discovery of information. It doesn’t contain any actual measurement data. The core layer also contains the controlled vocabularies for categorical measurements, and may be expended to contain other data types that need indexing. Additional indexes to speed discovery may periodically be built over the core, so that the operation of a metadata repository begins to resemble a web search crawler, except that it harvests observation series information.

Figure 2 shows the database diagram for the core profile. At the center of the layer is the `ODCore_SeriesCatalog` table. A catalog entry is defined by:

- Feature. The source location or spatial feature such as watershed for the data.

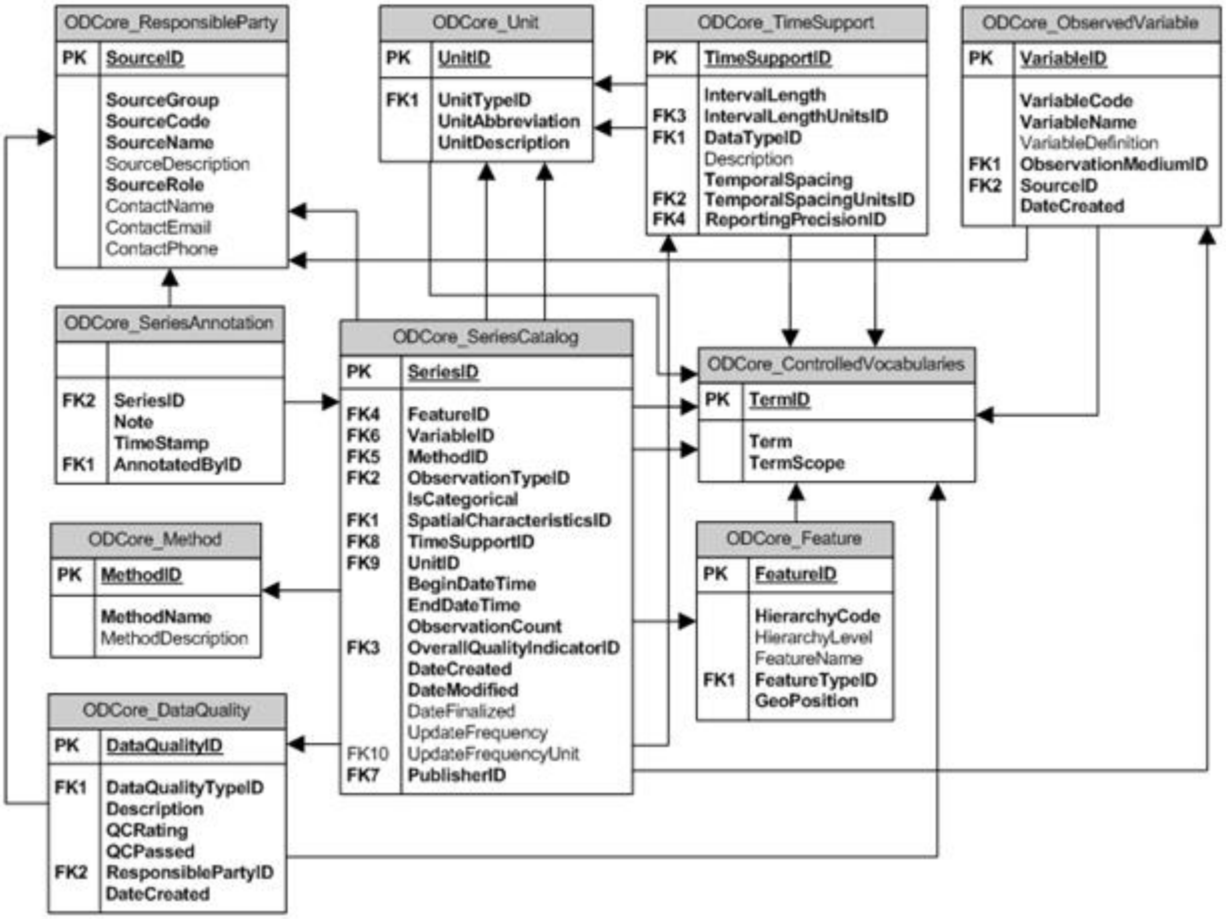


Figure 2 Core Layer

- Variable and Units. Variables may be gathered in many different units; unit conversion may occur dynamically when the data are displayed or statically when data are staged into the database or new variables are derived from existing data.
- Method Measurement method, particularly for laboratory samples.
- ObservationType. Indicates whether reported values are results of field observations, laboratory analyses, or numerical model runs.
- Spatial Characteristics. A given time series can be described as continuous coverages such as raster or single discrete value over the area of geographical feature to which it applies.
- Time Support. This describes the temporal characteristics of the measurements such as measurement frequency.
- Begin and end dates and times of the measurement series. If data collection is still on-going, the end time is the date & time of the most recent measurement.
- Observation count or number of data points.
- Overall quality. See section 6.
- Dates for series creation, last modification (including additions), and

finalization. A finalized series is frozen and will not be changed in any way.

- Update frequency and units
- Data publisher. Publishers may be government agencies such as the USGS (United States Geological Survey) or individual research scientist.

Controlled vocabularies are defined in `ODCore_ControlledVocabularies` table with TermScope attribute showing the context. For example, the term “Watershed” would have a TermScope of “FeatureType”.

2.2. Data Layer

Data layer is used for the storage of actual data i.e. observation results besides ancillary data including but not limited to the information about samples, laboratory analyses and data quality. Categorical observations utilize hierarchies in SQL Server for aggregations. For geographical features these aggregations involve spatial unions while for time series data, manipulation of the observation results is sufficient. Data storage is not limited to discrete coverages; continuous coverages such as gridded data or profile data are also supported.

Figure 3 shows the database diagram for the data profile. At the center of the data layer, is the `ODData_Observation` table. An observation is defined by:

- Series to which the observation belongs.
- Local date and time of measurement.
- Offset to UTC of that local date and time.
- Value. Actual observation data value.
- Category. Indicator of whether the value is numeric or categorical measurement.
- Vertical offset and offset reference.
- Censor. Provides identifiers such as “less than” and “present but not

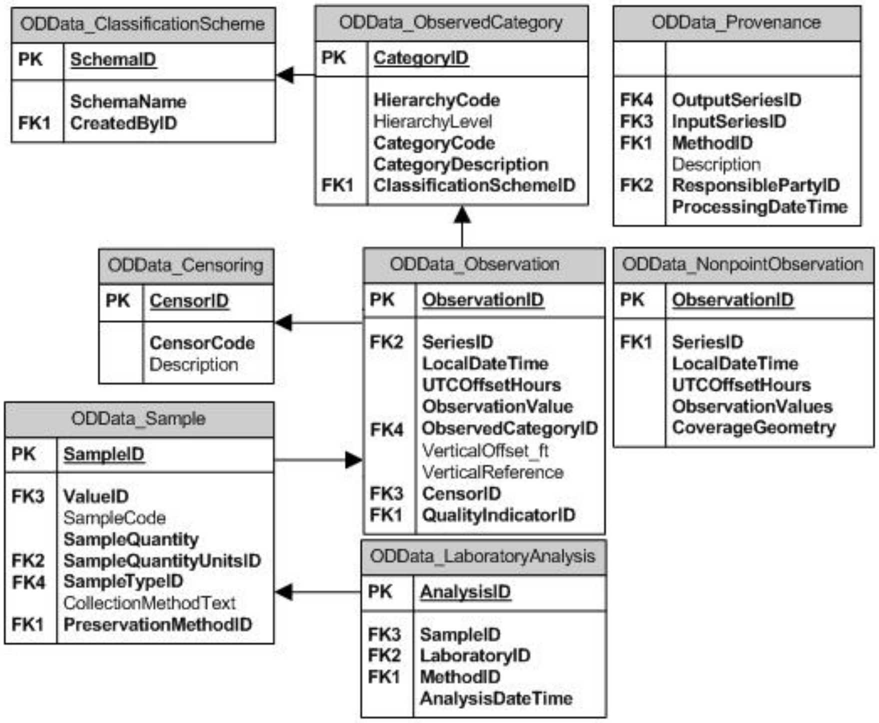


Figure 3 Data Layer

quantified”.
- Quality See Section 6.

Information regarding collected samples and laboratory analyses are stored in the `ODData_LaboratoryAnalysis` and `ODData_Sample` tables. Collection and processing details such as whether the same was obtained by a single “grab” or composite sample, sample quantity, and preservation methods such as the container, temperature, and chemical treatment are included here. Also included are contact information, sample number, analysis methods and analysis date for the processing laboratory.

2.3. Extensibility Layer

Extensibility layer enables users to declare their own properties that define geographic features or methods. One of the lessons learned from ODM experience was that every researcher has a different story to tell about their data. Some want to mention the casing of their well or how strong a wind their anemometer can stand; others want to report a standard deviation or a confidence interval for their measurement; still others want to report feature-specific attributes such as dam height or national aquifer code. These properties may have their own characteristics or sub-properties which the scientist may want to store. Scientists also want to annotate observation catalog entries that they generated or used in their research to point out errors or provide notes about processing. Extensibility layer also provides the underlying database structure for such metadata.

Figure 4 shows the database diagram for the extensibility profile. Each new or extended property is first defined in the `ODExtensibility_GenericProperty` table.

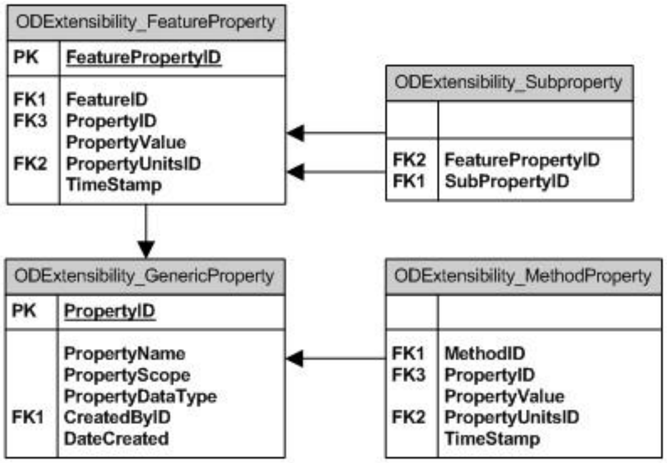


Figure 4 Extensibility Layer

Instances then can be used in `ODExtensibility_MethodProperty` or `ODExtensibility_FeatureProperty` with different units and timestamps. Section 6 explains how provenance is handled in the model.

2.4. Translation Layer

Translation layer serves as a dictionary which stores mappings between external and internal vocabularies. Sometimes researchers ingest data from other sources that don’t follow or are even in conflict with the controlled vocabularies in the database. Other times, researchers want to define their own vocabularies. For automating these processes a translation layer was necessary. Translation operation works on N-to-1 and 1-to-1 mappings.

Figure 5 shows the database diagram for the translation profile. Each vocabulary, or name space, is associated with a source publisher. Vocabulary publishers will tend to be the same as the data publisher for large agency data collections and smaller research collections, and may differ when the vocabulary is determined by a collaboration of individual researcher data publishers. Each vocabulary is tracked with a

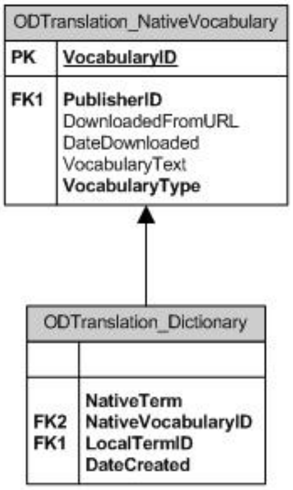


Figure 5 Translation Layer

download URL and date of last download. The `ODTranslation_Dictionary` spline table allows terms to be equated across vocabularies.

2.5. Collections Layer

Collections layer allows user to group data series they consider relevant and organize them for easy access. The data series can be grouped by original download, a specific processing algorithm, use in a publication or other 'folder'. Datasets can be accessed not only by the user who created them but also his/her colleagues.

Figure 6 shows the database diagram for the collections profile. The spline table `ODCollections_ObservationGroup` locates collection embers in the `ODCore_SeriesCatalog` table. Date stamps for collection creation, finalization and retirement are tracked. The accessibility role is currently reserved for future use and is intended to support user groups or role-based data access.

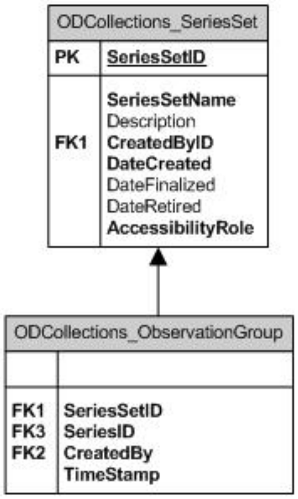


Figure 6 Collections Layer

3. Geospatial Characteristics

3.1. Gridded Data

Gridded data or rasters are data that have continuous coverage over an area as opposed to a single point. Rasters are part of the "data profile" and usually the result of model simulations, satellite or other remote sensing imagery, or dense sensor arrays.

Rasters are not natively supported by SQL Server 2008. We implement raster data with a user-defined type (UDT) [9] sometimes in the form of tiles depending on the file size. Each tile is georeferenced using a bounding polygon. Location of each grid cell relative to the bounding box is calculated using the row/column numbers and cell dimensions. Currently raster support is limited to quadrilateral grids.

3.2. Transect and Profile Data

Transects and profiles are paired values. These can be axes of a plot as in profiles or pairs of distances in x-y direction from a point of reference as in cross sections. Examples are temperature measurements over depth in a lake, river channel cross sections and soil horizon layers.

The time aspect is of less significance for these data. Time is not reported for each value in the set, instead a time representing the entire profile or transect is reported. Scientists are interested in the complete graph or cross section rather than parts of it unlike time series data.

Our schema also uses NonPointData UDT for storing data for transects and profiles. These data are georeferenced using the geography datatype and represented either by points (constant x,y varying z) or lines depending on the feature geometry. When a line is used, the ordering of points along the line indicates the ordering of the data points. Referential integrity rules are enforced programmatically.

Figure 7 shows some example use cases for NonPointData data type. The GRID on top is a 3x3 grid of 10 degrees resolution with columns and rows separated by spaces and commas respectively. The TRANSECT is a channel cross-section with each point presented by a pair of x-y coordinates separated by commas; the transect is defined with reference to the left channel bank. In both cases units are inherited from the corresponding "series catalog" entries.

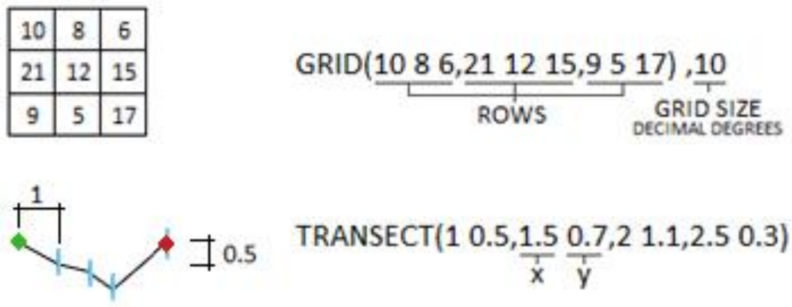
3.3 Point Data

In hydrologic science, point data is probably the most commonly used data type. These data are mostly products of stationary in-situ sensors or lab analysis of collected samples. In such cases observation locations are points that are independent of each other represented with their x,y,z coordinates. Each individual observation can be assigned a vertical offset. By default offsets are assumed from sensor's location, however an established vertical datum or other reference point (e.g. water surface, snow surface) can be defined explicitly.

Discrete coverages are not necessarily points or collections of points. For example average evaporation for a lake is georeferenced using the polygon that delineates the lake.

In case of a moving sensor, it is often useful to capture the information about the trajectory and relate the observations to one another as being the part of the same expedition. In the database trajectories are stored as polylines/linestrings in which the coordinates follow the order the data are collected.

Multiple sensor locations can be represented as collections of observation sites. HierarchyCode attribute in `ODCore_Feature` table is used to define parent-child relationships between sites and individual sensors. For example when working with a tower housing multiple instruments, scientists may prefer keeping separate records for each instrument in the database to be able to provide metadata on characteristics of each individual sensor. Similarly a scientist may have a number of sensors in close proximity to each other, or a number of random sampling points that are not visited more than once which he/she might want to represent as a single observation site/study area. This allows keeping record of



10	8	6
21	12	15
9	5	17

GRID(10 8 6,21 12 15,9 5 17) ,10
ROWS
GRID SIZE
DECIMAL DEGREES
1
0.5
TRANSECT(1 0.5,1.5 0.7,2 1.1,2.5 0.3)
X
Y

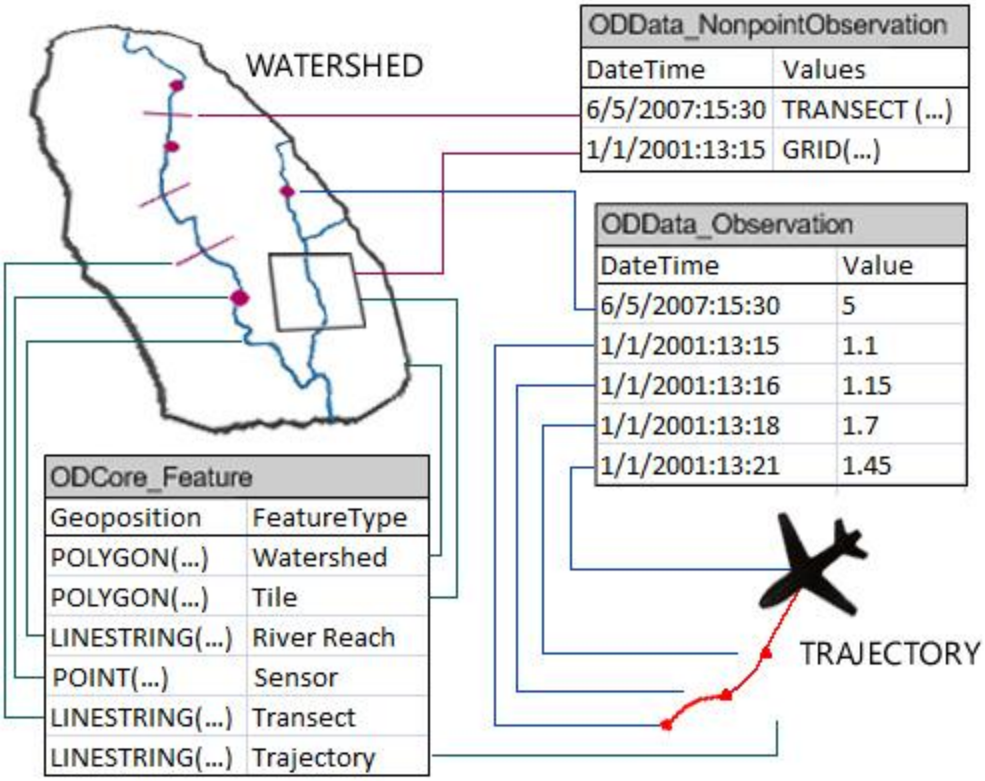
Figure 7 Creating Grids and Transects using NonPointData UDT

coordinates and characteristics of individual points without complicating the visualization and discovery of the sensors. Lastly, multiple trajectories can be generalized into a single, generalized trajectory. Servers that only serve as catalogs index these generalized geographic features.

Figure 8 shows how different spatial representations are handled in the database with two examples. The watershed and river provide the underlying geographical features which are also stored in the database. Transects are represented by red lines crossing the river, while sensors appear as points. Polylines are used to represent trajectories such as the plane flight path. Each measurement is a point along the polyline (i.e. LINESTRING) and listed in the order taken. For example, the first (earliest) measurement of a given time series is georeferenced by the first pair of coordinates of the polyline.

4. Temporal Characteristics

Data can be available in different temporal resolutions. Moreover for a given temporal granularity data can be reported in different ways such as incremental, maximum, or moving average. This information is captured in the `ODCore_TimeSupport` table. It is expected that aggregate measures will be reported with SQL dateTime precision. For example yearly maximum values should be reported on the date and time when the maximum value is observed, rather than just the year. This is only a recommendation. Some data sources may provide such data with coarser resolution or in some cases (e.g. yearly average) providing date and time with high precision may not be meaningful. For these special cases users are allowed to define the precision to be used with



ODData_NonpointObservation

DateTime	Values
6/5/2007:15:30	TRANSECT (...)
1/1/2001:13:15	GRID(...)

ODData_Observation

DateTime	Value
6/5/2007:15:30	5
1/1/2001:13:15	1.1
1/1/2001:13:16	1.15
1/1/2001:13:18	1.7
1/1/2001:13:21	1.45

ODCore_Feature

Geoposition	FeatureType
POLYGON(...)	Watershed
POLYGON(...)	Tile
LINESTRING(...)	River Reach
POINT(...)	Sensor
LINESTRING(...)	Transect
LINESTRING(...)	Trajectory

Figure 8 Use of discrete and continuous data types for different geographic features and observations

reporting.

5. Numeric vs Categorical Data

Both numeric and non-numeric data are supported by the data model. Non-numeric values are often ordinal measures like Beaufort wind scale (wind force) or nominal measures such as Anderson or International Geosphere-Biosphere Programme (IGBP) landuse/landcover classification. Each classification scheme is stored in the database with hierarchical relationships between categories allowing aggregations at different granularities.

For example a scientist looking for Anderson Level 1 (coarse) classification can derive the data from Level 2 (finer) data. As for numeric data, results can be associated with indicators such as ’less than‘, ’greater than‘, ’not detected‘ and ’not quantified‘ which may be necessary when an exact value is not reported due to detection limits or type of experimental procedure.

6. Data Quality and Provenance

Data quality information is captured at two levels: an overall data quality rating for the entire measurement series and/or separate ratings for individual measurements. In addition to a user-defined numeric rating, each assessment requires a Boolean pass/fail indicator to provide a common and simple data quality indicator across different datasets. Same observation or series of observations can have different data quality ratings depending on the aspect that is being evaluated e.g. temporal/spatial consistency, measurement accuracy. Hence textual descriptions as well as a data quality type indicator are included to provide the basis of the data quality assessment.

Data provenance or lineage is captured at multiple levels as well. From sample collection to preservation, analysis methods and through extensibility profile, intrinsic properties of these methods (e.g. accuracy, detection limits) provide an account of data lineage from the moment of collection to insertion into the database. Moreover data are created by processing other data in the database, the processing information can be captured with inputs, outputs, process description using `ODData_Provenance` table. Users can also employ annotations to convey more information about the data.

7. Reference Implementation

We have used the proposed data model to build a metadata using SQL Server 2008. We ingested geospatial data on hydrography, principal aquifers, dams, ecoregions and geology from USGS [10], EPA (Environmental Protection Agency) [11] and NCDC (National Climatic Data Center) [12] into the database. We also ingested catalog information for point data measurements from USGS, EPA.

SQL Server 2008 introduces two new datatypes, *geography* and *geometry* that allow vector data to be stored in the database. Considerable amount of geospatial data is freely available but mostly for consumption using GIS software; we needed to convert the available data through a series of conversions prior to database ingest. Since SQL Server 2008 doesn’t support datum conversions and re-projections natively, all the data was converted to the same datum (WGS84) in order to avoid errors and inaccuracies in geospatial operations due to inconsistent datums. Also, the GIS files need to be converted to the OGC Well-Known Text (WKT) or Well-Known Binary (WKB) formats to be inserted into the database [13]. All necessary conversions were done using SAFE Software's FME Workbench geospatial ETL[1] tool [14]. Properties of different features were handled using the `GenericProperty` and `FeatureProperty` tables in `Extensibility` profile.

Stream networks were created based on National Hydgrography Dataset [15] using hierarchyid datatype. Hierarchy requires a given child to have one and only one parent. Process of network building can start from sinks or sources. Regardless of the direction the hierarchies are built, a simple hierarchy cannot be relied on since rivers can split and converge. Solutions that involve recursive T-SQL queries or CTE[2] also fail since rivers may diverge at a point then converge further downstream which causes infinite loops using these approaches. The key insight to solving this problem is that the order the streams are digitized indicates the direction of flow. To create a hierarchy SQL Server's STEndPoint() and STStartPoint() methods can be used to link the last point of the polyline that represents a stream reach with the first point of the polyline that represents the stream reach downstream of it. When multiple parents are required in the hierarchy, a junction is created at the point of convergence which serves as the endpoint of the additional stream reaches. Navigation of the network is handled in the query level during which a stream reach ending with a junction indicates that at the same exact geographical position there exists a stream reach that continues upstream. In the current implementation hierarchies start from sinks since rivers tend to converge more than they diverge thus this approach reduces the number of necessary junctions.

Our catalog contains data from approximately 1.6 million observation sites operated by agencies such as USGS and EPA. Our database contains metadata for 9.3 million data series corresponding to more than 358 million measurements residing in the cloud.

Our catalog makes use of Core, Extensibility and Translation layers and underlies the SciScope data discovery tool (www.sciscope.org) [16].

8. Conclusions and Future Directions

Environmental data is rich in data types, requirements for time reporting and variable and location naming conventions. Data from human field observations, sensors, satellites and simulations are often combined. Science variables are often derived from the original data; data cleaning is ongoing. Science data analysis often includes aggregations over space and time.

Our early experience with this kind of data focused on sensor data and sample data sent to laboratories. In the process, we learned the importance of spatial data types such as rasters. We also experienced the naming, units conversion, and provenance problems associated with trying to use data gathered from a number of different producers including government agencies, science or other collaborations, and individual researchers. Attempting to use a "one size fits all" schema simply won't work.

Our proposed design for ODM v2 adds spatial support, new data types, name space translation, formalizes extensibility, and adds minimal provenance and annotations. While

[1] Extract, Transform, Load

[2] Common Table Expressions

still a work in progress, we believe it is an important step forward to digital watersheds and other cyber-laboratories.

We continue to operate our reference catalog server and will be adding metadata for other environmental data sources as they become available. We are also working on converting our existing BWC California digital watershed to the proposed data model as a reference data server implementation. That watershed includes hydrologic data accumulated from USGS, NOAA, NMFS, and other agencies around the state. That watershed has approximately 90M data values as well as channel cross sections, sediment grain size distributions and other non-time series measurements which we believe will further validate the extensions for science use.

References

[1] B. Plale, D. Gannon, *et. al.* "CASA and LEAD: Adaptive Cyberinfrastructure for Real-Time Multiscale Weather Forecasting", Computer, Vol 39, issue 11, November 2006, pp. 56 – 64.

[2] Aman Kansal, Suman Nath, Jie Liu, and Feng Zhao, "SenseWeb: An Infrastructure for Shared Sensing," IEEE Multimedia. Vol. 14, No. 4, pp. 8-13, October-December 2007.

[3] Maidment, D. (Ed.), 2005. Consortium of Universities for the Advancement of Hydrologic Science Inc. Hydrologic Information System Status Report, 224 pp.

[4] NSF cyber observatories, http://www.cyberobservatories.net/

[5] WATer and Environmental Research Systems Network, http://www.watersnet.org/

[6] Berkeley Water Center, http://bwc.berkeley.edu/

[7] Beran B., Piasecki M., 2007. HYDROSEEK: A Search Engine for Hydrologists, Geoinformatics 2007, May 17–18, 2007, San Diego, CA.

[8] Whitenack T., Zaslavsky I., Valentine D., Djokic D., 2007. Data Access System for Hydrology, American Geophysical Union Fall Meeting, December 10-14, 2007, San Francisco, CA.

[9] CLR User-Defined Types http://msdn.microsoft.com/en-us/library/ms131120.aspx.

[10] United States Geological Survey, http://www.usgs.gov and http://waterdata.usgs.gov/nwis .

[11] Environmental Protection Agency, http://www.epa.gov .

[12] National Climatic Data Center, http://www.ncdc.noaa.gov/.

[13] Herring J. R., (Ed), 2006. Implementation Specification for Geographic information - Simple feature access - Part 1: Common architecture, Open Geospatial Consortium Inc. , 95 pp.

[14] SAFE Software FME Workbench, http://www.safe.com/products/.

[15] USGS National Hydrography Dataset http://nhd.usgs.gov/.

[16] Beran B., 2008. SciScope: Geoscience Data Discovery and Visualization using Virtual Earth, Virtual Earth™ and Location Summit 2008, April 30-May 1, 2008, Redmond, WA.