Use Cases to Test OGC O&M Profile

Background

As we develop experience in having scientists both use CUAHSI HIS and publish data in CUAHSI HIS, we are finding some difficulties arising in interpreting the data when a simple location/time/parameter name isn't adequate to describe the data. Scientists need to know more about the *context* of the data collection—this could include the intent of the scientists who collected the data (what does this measurement represent to the scientist?) and the environmental setting of the measurement (what is the "site type"?).

This issue has not been very fully explored with the HIS project because we have considered the stream gage as the archetype of our measurement location. In this regard, the most important underlying independent variables describing the gage data are i) location, ii) sampling time. A gage on a river provides a great deal of context by its location—one can readily understand the position of the gage in a river network by looking at a map, streamflow data is understood to reflect hydrologic conditions at that point of the river and to convey information about the hydrologic status of the watershed it drains. The terrestrial setting is defined by properties of that watershed, such as land cover.

However, as we extend to different measurement types, the structure and efficient transmission of information about the environmental context and interpretation of the data differ for each type of data. The attached "flower" diagram from the Critical Zone Exploration Network (www.czen.org) is a useful depiction of the range of data types that exist to describe the critical zone.

The intent of this discussion piece is to pose some use cases to see how the Open Geospatial Consortium's Observations and Measurement (O&M) profile can be adapted to capture this broader contextual information about data.

OGC Observations and Measurement Model

Reference: https://www.seegrid.csiro.au/wiki/AppSchemas/ObservationsAndSampling

Here, the O&M model is described by quoting parts of the reference. I have focused on how this information model is applied to the interpretation of data rather than formal details of its representation.

Starting with a simple definition of the O&M information model:

> An **Observation** is an action whose **result** is an estimate of the value of some **property** of the **feature-of-interest**, obtained using a specified **procedure** [at a specified **time** by a certain **party**.]

This is the simple abstract information model. For the purposes of this discussion, I have added the concept of time and party because time series (the default data type for CUAHSI HIS) requires time and

the concept of the collecting entity ("party") is required for data assemblage from multiple groups. Note that time is central to our central data type—the time series—but that for other environmental measurements, especially solid phase chemistry, time of sample collection may be irrelevant because of the long time periods involved, as indicated on the CZEN Flower Diagram.

Let's start with a default use case. (All this information is fictional.) Consider a measurement of the stage of the Colorado River at Austin, Texas which is 2.2 ft at 13:00 GMT on 9 August 2011 as determined by the USGS. Some of these words are quite straightforward:

- **Result** is the measurement value (2.2 feet)
- **Procedure** is the measurement method. ("stilling well using a pressure transducer [manufactured by xxx, model number yyy] logged by a Campbell data logger and transmitted to the GOES satellite to USGS ADAPS system where quality controlled using USGS method ZZZZ" or perhaps just a "USGS certified method.")
- **Property** is the measured property ("parameter" in USGS parlance).
- **Time** is 13:00 GMT on 9 August 2011
- **Party** is USGS.

In this example, the "result" of stage may be thought not to require interpretation, because we are assuming that the voltage signal from the pressure transducer has been properly interpreted. If the example had been "discharge" rather that stage, this value results from a rating curve that must be developed and maintained by the operating party. If we trust the USGS to maintain that curve, perhaps we accept the discharge as valid. The point here is that this model is not making a distinction between a direct measurement and an interpretation of that measurement because this distinction is implicitly viewed as one of degree. As Simon Cox points out, even temperature measured with a mercury thermometer involves interpretation (i.e., converting the length of the column of mercury to temperature).

The more subtle phrase in the definition is "property of interest." In this example, is the property of interest the gaging station (i.e., the river crossection), the river reach (i.e. that portion of the river defined, for example, between two confluences), the entire river, or the watershed drained by the river?

Before we decide an answer, we need to introduce the concept of *sampling*. From the same reference, sampling is required in either of 2 cases:

a. the feature is inaccessible (e.g. concealed, or too large for exhaustive observation)
b. the properties are not directly observable (e.g. the feature is remote, or for other reasons does not provides a direct physical signal)

In this case, the feature of interest as a *sampling feature* that is accessible and provides a direct signal. Sampling features have a shape (point, curve, surface, or volume). A *specimen* (i.e., sample) can be taken of a sampling feature for *ex situ* analysis. A single specimen can result in multiple observations (e.g., measuring many analytes on a sample).

These concepts of *feature of interest, sampling feature,* and *specimen* provide a rich information model in which to capture the environmental context of an observation.

Simon Cox elaborates on *sampling features:*

> *"*There are interesting relationships between sampling features and other features. In particular:
>
> - every sampling feature exists because of an *intention* to sample or represent one or more domain features [that is, the feature of interest]. For example
> - an **ObservationWell** (a kind of SamplingCurve) samples one or more **Aquifers**;
> - a **RockSample** (a kind of Specimen) samples a **GeologicUnit**;
> - an **Outcrop** (a kind of SamplingPoint) samples a **GeologicStructure** and one or more **GeologicUnits**;
> - a **Scene** [i.e., remote sensing scene] (a kind of SamplingSurface) samples a **Landscape**
> - a sampling feature exists because observations have been or will be made which utilize it. For example
> - an ObservationWell may carry a set of Logs (observations whose result is a 1-D Coverage);
> - a RockSample may carry a set of Assay measurements and Geochronology measurements
> - samples are commonly related to other samples, through sub-sampling, as part of an array, etc. For example
> - Intervals (SamplingCurves) and Specimens may be contained within, or retrieved from, an ObservationWell;
> - a RockSample may yield a set of Splits or Separates (sub-samples);
> - a Specimen may be taken from an Outcrop
>
> For example, a specimen must be tied to the domain feature (e.g. an organism, a material etc, as its *sampledFeature*), but may also be associated with a Sampling Point or interval, etc, if the details of the sampling location within the domain feature are of interest
> Note that the sampledFeature is expected to be a 'domain feature' - i.e. a real-world feature that is not a sampling feature or observation.
>
> In fact the relationship between observations, sampling features and domain features is subtle, and *data providers may make different choices about how much information about the sampling regime to provide to the data consumer*. [emphasis added] For example, a provider may choose to indicate

the domain feature as the observation feature of interest, and bundle the description of the sampling feature with the observation process as a 'protocol'."

In the default use case of a stage observation on the Colorado River at Austin, Texas, what is the feature of interest? It could be viewed as a direct observation of the river cross-section (a real-world, domain feature), or it could be viewed as a sampling feature of the Colorado River. In the former case, the cross-section is "exhaustively" sampled by virtue of either a natural or constructed control feature, the gaging station is chosen so that the stilling well provides an accurate measure of water level across the entire crossection (e.g., the river is not braided). The latter interpretation of the feature of interest as the river may be more appropriate if the intent is to measure the stage at every point along the Colorado River but the river is too large to sample exhaustively. The cross-section at Austin is a sample of the Colorado River (the feature of interest). The latter interpretation makes sense in the context of flooding, for example, where the interest really is in the water elevation throughout the river. One interpretation is not superior to any other because it has to do with how the data are to be interpreted. I believe our implicit intent in HIS has been that the feature of interest is the river cross-section.

However, as we move away from an *in situ* measurement at a river gage, this specification becomes more important.

Case #1. Spatial Averaging

Use Case #1a: Soil Moisture

A time series of soil moisture is measured at site 325 at a depth of 10 cm at Panola Mountain watershed using a Decagon Probe. The soil moisture was 20% on January 15, 2011 at 7:30 am. There are additional probes located at 30, 50 and 90 cm of depth that are located as near to site 325 as possible given the logistics of installation.

In this case, the sensor is influenced by a few cubic decimeters of soil surrounding it. That portion of the soil is the "sampling feature" which is simply the spatial support of the sensor. However the intent in placing the probes is not just to measure that sampling feature. Rather, I interpret that number to be the value of the soil moisture between 0 and 20 cm for an extensive area (in this case, the flood plain between two tributaries at Panola). That interpretation is based upon a level topography, similar vegetation cover, and a uniform soil type over that spatial extent. The geovolume is the feature of interest for this measurement.

This use case is a common one for many sensors deployed in the terrestrial environment. I believe that the scientist should record the intent of the measurement (in this case the geovolume defined as the soil layer between 0 and 20 cm with an extent of between two tributaries in the floodplain) as well as the sampling feature (the location of the probe and, by virtue of the probe's design, the spatial support of the sensor). This measurement may, in fact, not be representative of the geovolume, but that is the data provider's interpretation of the data. As such, he/she is more qualified to make that interpretation than anyone else. In addition to recording the geovolume, additional metadata should be recorded

about the basis for this interpretation (i.e., the topography, etc. listed above). The metadata could also include synoptic sampling that confirms the representativeness of the sensor.

Use Case #1b. Rain Gage Network

A network of 6 tipping bucket rain gages calibrated to record every 1 mm of rain are located throughout Panola Mountain watershed. In addition a weighing bucket rain gage serves as back-up when rainfall intensity exceeds the tipping bucket capacity. Both the tipping bucket and the weighing bucket have an opening of 400 mm^2 and are surrounded by wind shields. Each has been placed in a clearing where there is an opening of at least +/- 45 degrees, as dictated by standard protocols. Values are combined using a Thiessen polygon approach.

The intent of this rain gage network is to provide a basin-wide estimate of precipitation—i.e., that is the feature of interest. The sampling feature is a tiny portion of the entire basin area, but this is a well-studied problem.

Clearly, the data of interest to most people using CUAHSI HIS is the basin-average precipitation, not the area sampled by individual rain gages (unless someone is studying rainfall fields). Presumably both the individual records and the basin-wide average should be published, but the question is how the method of the basin-wide average should be captured and transmitted to the user of the data. The existing CUAHSI Observations Data Model (ODM) contains appropriate fields (including "derived from…") but a best practice should be developed to ensure consistent reporting of these spatial averages.

Use Case #1c. Snow Course data

Three snow courses are measured at Emerald Lake Watershed once per month for snow water equivalent. Data from the snow courses are combined based on elevation and aspect to derive the basin-wide snow water equivalent stored in the snowpack.

This use case is essentially the same as the rain gage network, but with fewer well accepted protocols. Again the question is how to capture the metadata supporting the basin-wide average.

Use Case #1d. Sapflow measurements

Water uptake in a mature oak tree that extends into the main forest canopy is measured using heat dissipation. The oak tree is located on a hollow in a hillside at Panola Mountain at latitude xxxx and longitude yyyy.

For a hydrologist, the intent of the measurement is to estimate basin-wide transpiration and, perhaps, this oak tree is part of a stratified sampling strategy determined by species (oak), landscape position (a hollow) and canopy position (overstory). Then the feature of interest is basin-wide transpiration and the oak is a sampling feature. How do we capture the stratified sampling design in the metadata?

The question also arises how an ecologist or a tree physiologist would look at this data. What other metadata should be captured about the tree?

Use Case #1e. Water level in a well

(a) A well screened in aquifer qqqq has a water level of 54 feet below land surface on September 3, 2011 at 14:50.
(b) An open borehole drilled to a depth of 250 feet that intercepts a number of geologic formations has a water level of 24 feet below land surface on July 23, 2011.

What is the feature of interest in each of these cases? In case (a), it seems to be aquifer qqqq. In base (b), we can't attribute that water level to any one formation, yet there is information here. How should that be represented?

Use Case 2. *Ex situ* Analysis

The CUAHSI Observations Data Model and the WaterML 2.0 transmission language were largely designed for sensor data, that is *in situ* data. However, ODM has been used for simple aquatic chemistry data derived from samples analyzed in a laboratory (that is, *ex situ* data). EarthChem provides a far more extensive set of organized metadata for *ex situ* analyses. When should ODM and when should EarthChem be used for these data?

Use Case 2a. DOC concentration

The dissolved organic chemistry of a sample collected at site 103 at Panola Mountain on June 2, 1996 at 13:05 has a concentration of 6 mg/L.

Feature of interest: The creek or the cross-section of the creek.

Data base consideration. The parameter name "dissolved organic carbon" conveys a lot of information about sampling method. The sample has been filtered ("dissolved") and "organic carbon" implies that a complete digestion of carbon has been performed after removing the inorganic carbon from the sample (e.g., by acidification and sparging the sample). So, DOC data placed in ODM is completely interpretable because the parameter name captures sufficient information about the method.

Use Case 2b. Suspended sediment concentration

The suspended sediment concentration of a sample collected at site 103 at Panola Mountain on June 2, 1996 at 13:05 has a concentration of 30 mg/L.

Feature of interest: Seems like it has to be the cross-section as suspended sediment concentration is highly variable with zones of mobilization and deposition, but this is really an issue of how far one could extrapolate and what we mean by a 'sample' of a river when the property is not easily extrapolated beyond the measuring point.

Database: ODM is again adequate because the method is sufficiently well defined by the parameter name to make the data interpretable.

Use Case 2c. Chemistry of suspended sediment

The silt fraction of the suspended sediment of a sample collected at site 103 at Panola Mountain on June 2, 1996 at 13:05 has a concentration of 50 mg Ca per kg of sediment (or yy mg/L in the sample) as determined by ICP/MS and using a partial digestion to estimate the 'readily available' calcium.

Feature of interest: The suspended sediment at this river crossection?

Database: EarthChem must be used because "calcium" no longer defines the measurement. Analytical methods must be explicitly documented.

Use Case 2d. Foliar chemistry

The calcium concentration of a needle from a hemlock (*Tsuga canadensis)* at the Harvard Forest is 40 mg/Kg on March 15, 2005 at 14:31.

Feature of interest: The narrowest interpretation is that the feature of interest is this particular hemlock tree. What metadata should be captured about it (age, health, canopy position)? Alternately, is this a sample of all hemlock trees in the Harvard Forest? Perhaps the distinction is whether an average value is published in the data base (e.g., the foliar content of calcium in a comprehensive calcium budget of Harvard forest) which would then be documented through individual samples (such as the one mentioned here), a sampling strategy, and an algorithm for space/time averaging of samples.

Database: Could EarthChem handle this kind of sample?

I hope these use cases can be added to by other communities to help inform our discussion of data interoperability in EarthCube.