- Use relevant schema.org terms in describing variables whenever possible
- Determine what new extensions to schema.org may be necessary to effectively describe variables
- There are some well-established and maintained ontologies available to clarify the contents of a variable description, these need to be accessible online to use. Developing and adopting these is out of scope for schema.org.

TIER 1

The simplest thing is for **variableMeasured/PropertyValue** to have a **propertyID** that is a URL (http URI) that points to a dereferenceable term in an Ontology that represents the variable, e.g. **http://purl.obolibrary.org/obo/ENVO_04000002** .

That term might have a rich representation, accessible by dereferencing the URI, that further axiomatizes the phenomenon, e.g. as a "Feature of Interest: **sea surface"** and "Observed Property: **temperature**", etc, with links to ontologies like EnvO or SWEET.

This value in an instance documents will be a (potentially *opaque)* URL; it would be up to the client to extract useful information from the source semantic resource to better understand what the PropertyValue actually represents, via its **rdfs:label**, **skos:definition**, **skos:altLabel**, etc. Potential solution: conventions for how to represent such descriptions and/or definitions in schema.org instance documents.

TIER 2

The next "tier" might be to include some of the most useful pieces of information found in the **rdfs:label, rdfs:comment**, (or some SKOS field etc) associated with a **PropertyValue/propertyID/URL** in schema.org instances. For this, one could use the "**name**" and "**description**" properties of the "PropertyValue", with **text** extracted from the referent ontology. Thus, one would see that some "**variableMeasured**" was of a "sea surface temperature", ideally with some description or definition of the variable.

schema.org has some additional "pending" elements that use "**Defined**" in their labels in a fairly loose way semantically speaking:

DefinedTermSet can be used to identify an Ontology or controlled vocabulary using a **URL** and **name**, that might pertain to ALL the variableMeasured types in some Dataset, or for each PropertyValue.

DefinedTerm identifies a concept via a URL, and can also have a name, alternate name, identifier, comment, same as, etc. Unfortunately the proposed list of properties that can take DefinedTerm as a value do not include any properties that apply to **PropertyValue**, so we'd have to extend Schema.org.

termCode is a property of **DefinedTerm**, and only allows a **TEXT** value. It be an abbreviation or the hash-suffix that identifies the Term in the DefinedTermSet (note that deconstructing Identifiers in this way, only to have to reconstruct a URL that is dereferenceable, is problematic!)

TIER 3

Any "**variableMeasured**" is accompanied by properties that conform to a formal Observations/Measurements (OM) model. This could be done by extending the schema.org model for the **PropertyValue** object, or defining a new property value type object consistent with the OM

model used, and add that to the range of **variableMeasured**. Either approach is viable given the open-world nature of RDF, but for interoperability, a minimal amount of modification of existing schema.org elements is recommended.

This would involve extensions such as "Entity of Interest" and "Characteristic Measured", that may also need to have their **URLs, names,** and **descriptions,** etc. exposed via schema.org and JSON-LD

TIER4

Variables with object values. In all above there is an assumption that the "variableMeasured" pertains to some set of actual measurements within a Dataset, e.g. all the values found in a column of some Table. Some "datasets" will likely want to describe their "variableMeasured" at a more atomic level, which is possible given the above approaches if it is some "BLOB" (e.g. a SHAPEFILE) that has some well-defined metadata terms that might be construed as "variableMeasured".

A harder Use Case is when we have a more complex data type-- e.g. a multi-dimensional table, hierarchical data structure, etc. unless we have some "index-identifier" that can unambiguously point to individual values within some data object (e.g. in a Matrix, it could be [2,3] to indicate entry in second row, third column, but in a Table such row and column ordering is often not rigidly constrained.

What is needed for the more complex PropertyValue value types is a richer 'dataType' property that could reference standard (e.g. xsd:) datatypes, or a datatype registry (a la RDA Data Type Registries working group) entry via a URI.

Commented [SR1]: I interpret this to be about a variable whose value is an object of some sort, in this case a shape file. I think variables whose values are images, audio or video recordings, or maybe even big NetCDF objects with a model output.

Commented [SR2]: I'd lump this in as an extension of the above types, like a netCDF or shape file, these would be measuredVariable values that are themselves data objects.

Additional considerations:

measurementTechnique. Redux of documentation at https://schema.org/measurementTechnique

> *A technique or technology used … for measuring the corresponding variable(s). Not intended as a full representation of measurement, but rather as a high level summary for dataset discovery.*
>
> *For example, if variableMeasured is: molecule concentration, measurementTechnique could be: "mass spectrometry" or "nmr spectroscopy" or "colorimetry" or "immunofluorescence". If the variableMeasured is "depression rating", the measurementTechnique could be "Zung Scale" or "HAM-D" or "Beck Depression Inventory".*

Value is Text or URL

Some way to specify units of measure as appropriate for PropertyValue specification. **unitCode**, and **unitText** are additional existing schema.org properties that we should recommend for clarifying the contents of a "**variableMeasured**". Again, use of **URL** and **name** will be invaluable in unambiguously identifying what these are, although "**unitText**" only admits of a TEXT descriptor so would often be the "name" associated with the **unitCode** (**URL**). [Sci on SDO should recommend a vocabulary for unitCode URIs]

Precision is a term that appears to be missing from the schema.org vocabulary. We might want to recommend it for inclusion as a new property of a "**variableMeasured**"

Summary thoughts

Goal:

Describe datasets for discovery, evaluation, and access

For discovery,

- basic dataset name, description and keywords are a good start.
- Goal is to add information about the variables that are specified for data items in the dataset to support deeper search, tier 1 is variable name and a URI referencing some authority.

For evaluation

Need to know something about

- Measurement technique
- Data quality (precision, accuracy, validation procedures…)
- Value range in data
- Units of measure
- Observation context -- many datasets can benefit from some environmental contextualization. Some relevant properties include:
    - **biome** (arctic tundra) where the dataset was collected
    - **habitat** (thermokarst) where the dataset was collected;
    - **environmental feature** that was sampled (thaw lake) [sampling feature and feature of interest?]
    - **environmental material** that was sampled (talik).

These might be Properties that apply to the entire **Dataset**, or to a specific "**variableMeasured**" within a Dataset as environmental feature and environmental materials may vary across measurements within a dataset.

The properties specifying context should have values of (at least?) **name** and **URL**

- Value Types—data types including e.g. simple literals (integer, decimal, float, text), links, structured objects, binary objects (image, audio, video)