

<MMI /> Workshop

Advancing Domain Vocabularies

Guidance: The Meaning of “Same As”

4 August 2005

Introduction

For each of the vocabulary terms you will be comparing, you will be asked to decide their relationship to each other—specifically, whether one is the same as, broader than, or narrower than the other (or none of the above, or even if some other relationship should be created).

How should you make this decision? What characteristics or attributes should you consider? Possible attributes could include the measurement type, the thing being measured, the way or place a measurement is made, and so on. This could be quite complicated to figure out.

In an attempt to standardize the mapping results, we have developed some recommended criteria. If you find these criteria are not appropriate, please consult with the workshop organizers as part of deciding on alternative criteria. We may be able to apply your situation to the other teams.

The details of our analysis may be found on the reverse side of this document. We encourage interested participants to read and comment on this material as part of their workshop results.

Guidance

For most of the domains in this workshop, those focused on scientifically measured or composed variables, we recommend using a the **measurable property**, together with the **domain** (or **measurable particular**) being measured, to decide if two entities should be related.

Examples of measurable properties include: temperature, length, volume, speed, and the many other entities referenced by the Systeme Internationale for units. Measurable properties include derived measurements, using computed values to classify a thing, for example a *dominant* biota.

Domains are the substances being measured. At the simplest level, this can be water, air, soil, and so on. More controlled and detailed vocabularies can be established; an example of these may look like: hydrosphere-ocean-midwater. The GCMD vocabulary list is a useful reference.

We believe the maximum utility will be obtained by combining a fundamental measurable property with as simple a domain reference as is coherent. For example, “water temperature” is more general categorization, and therefore more often usable, than “sea surface temperature”.

Analysis: The Meaning of Same As

Options and Alternatives

There are many discriminating components that made up a description for a data value:

- measurement property (length, temperature, voltage, ...)
- domain (water, air, ocean, ...)
- units (and scale)
- depth relative to ocean surface; location (latitude, longitude)
- date/time
- sensor type
- platform type
- specific platform or observing campaign
- post-processing discriminators (raw, QC, avg)
- other discriminators, like frequency, cardinality, interval, bin size...
- BODC's "measurement process" is probably an additional discriminator

These discriminators could be used to distinguish between items in any of the following ways:

- A) However the people in the room feel about two items defines whether they are different
- B) Any difference in any discriminator makes the item different
- C) Only differences in selected discriminators make two items different
- D) Only one discriminator applies to decide if two items are different
- E) Each discriminator is mapped separately, so queries can be composed for each discriminator separately

(A) is the most intuitive, but also the most useless, since there are no rules or repeatability to the process. (E) is the most precise and computationally satisfying, but is very difficult to implement, and still is not a fully constrained solution.

(B) is intuitive and appealing in certain contexts – for example, a different location or measurement technique feels like a different kind of item. But this technique breaks down in many real-world cases. For example, when a moving sensor continuously measures values, as on an AUV or tow-yo, every measurement would be different than every other measurement, which is surely not the most desirable outcome.

(D) is minimally ambiguous, especially if the measurement property is chosen as the one discriminator. So it perhaps holds the most computational promise. But it is socially unsatisfying—the result is too trivial and counterintuitive to most users, who want "same as" to be more scientifically meaningful.

This leaves (C), an agreed set of discriminators. The most realistic candidates, based on a little experience with about 20 significant data sets and repositories, are measurement property and domain. One can imagine using these two together as a scientifically useful basis for determining relationships, despite some likely ambiguity. (Is water temp different than ocean temp, if the first is in the estuary and the second in open ocean?) Since it is generally the case anyway that ambiguity is unavoidable, and judgment will take over, we choose C as our recommendation.

Notes

The mapping of **measurable particulars**, or physical entities—sensors, rocks, geologic features—depends heavily on the uses for the mapping. The teams will have to decide what level of detail and precision will produce the most useful relationships. It will be very useful to capture the nature and results of those discussions.

The **measurable property** is easily defined as the collection of the component definitions:

- measurable: capable of being measured
- measured: assigned numbers to phenomena according to a rule
- property: basic or essential attribute shared by all members of the class

Other examples of detailed hierarchical domains: atmosphere.stratosphere, lithosphere.seafloor, and observatory.platform.-instrument.sensor. Distinctions between one detailed domain and another are often too subtle to represent a real difference.

References

Ontomet: Ontology Metadata Framework. Luis Bermudez, Dec 2004. <http://dspace.library.drexel.edu/handle/1860/376>
See particularly Chapter 10 for an ontology used as the basis for the boldface terms here.