Vocabulary Innovations Introduction

Roy Lowry

BODC Emeritus Fellow

50

The Problem

The $64,000 question for data aggregation interoperability:

- Did you measure the same thing as me?
  - *'The same thing' is a context dependant moveable feast*
    - Is a measurement on filtered water the same as that measurement on unfiltered water?
    - Is a CTD salinity the same thing as a salinity by silver nitrate titration?
- What we really want to know is what measurements may sensibly be incorporated into a given data product

The Problem

We ‘know’ what was measured because the measurer provides the measurement with a label

- BUT scientists’ measurement labels can be parochial
  - *Temperature:*
    - Of ocean or atmosphere??
  - *20-carbon alkane:*
    - n-icosane, one of the other 366318 isomers or an isomer mixture??
- Organic chemical and biological taxa names are often mis-spelled

A Solution

Build a collection of ‘what was measured’ labels and give each label a unique identifier

Scientists tag their data with the label identifier that fully describes their measurement

Build a collection of data aggregations labels and give each a unique identifier

Engage science domain experts to define each data aggregation label as a list of ‘what was measured’ labels

Problems with the Solution

As the number of measurements and users grow some problems emerge

- Inconsistent phraseology, spelling and word ordering
  - *Technological solution: semantic modelling and integration of standards (WoRMS, CAS, ChEBI)*
- Ensuring all users are using the same list of measurement labels
  - *Technological solution: vocabulary server*
- Inability to find the relevant measurement label
  - *Technological solution: vocabulary search tool*
- No label exists for a measurement
  - *Technological solution: vocabulary builder tool*