LTER Data Set Attribute Standardization: First Steps and Discussion

LTER IMC Virtual Watercooler, 1-2 Nov 2010

Corinna Gries (NTL)

Wade Sheldon (GCE)

Why Standardize Attributes?

- Attribute labels, descriptions extremely important for data discovery, evaluation (just below title, abstract)
- Non-standardized attributes make everything harder at the network level
  - Cross-site data search and evaluation
  - Cross-site data integration/synthesis
  - Search interface development
  - Workflow development and re-use (e.g. EcoTrends)
  - Semantic mapping to ontologies
- Major criticism of LTER IM (both insiders and outsiders)
  - Heightens contrast to EONs (not in a good way)
  - Makes us look disorganized
  - Highlighted in Decadal Plan, NEON “legacy data” prospectus

Levels of Standardization

- Standardization can occur at multiple levels
  - Concept only: controlled vocabulary for labels (e.g. Temperature)
  - Concept + context: controlled labels for attribute and measurement medium/method and scale (e.g. Daily_Mean_Temperature_Water)
  - Concept + context + methodology: (e.g. Daily_Mean_Temperature_Water_ThermistorString)
  - Ontology (or ontology reference)
- Attributes can be standardized as
  - Labels (e.g. ClimDB)
  - Codes (USGS NWIS, STORET)
  - Labels + codes (i.e. method code in separate field)

Examples from Other Programs

- USGS NWIS
  - Use modified USEPA STORET Codes
    - 5 digit numeric code for each measured parameter
    - Code represents parameter, medium, method, units, etc.
    - NWIS appends 5 digit code for aggregate measures (e.g. 00001 = maximum, 00002 = minimum, ...)
    - Example:
      00094_00001 = Maximum specific conductance, water, unfiltered, field, microsiemens per centimeter at 25 degrees Celsius
  - Strengths
    - Attribute metadata easy to maintain (once mapped)
    - Reduces need for extensive methodology metadata
  - Weaknesses
    - Requires resolution service to explain codes
    - Cryptic – end-users generally need to re-label for use
    - Varying level of specificity (fragmentation over time)
    - Registration process slow, leads to “provisional” codes, force-fits

Examples from Other Programs

- CUAHSI
  - Standardized Attribute Name, e.g. Temperature
  - Standardized units
  - Standardized sample medium
  - Attribute ID not standardized and combines:
    - Attribute Name, unit, sample medium (e.g. surface water or air)
  - Variable methods
    - Variable measurement frequency
  - Strength: will find all ‘Temperature’ data, and easily all ‘surface water temperature’
  - Weakness: for evaluation and use methods have to be resolved.

Potential First Steps in LTER

- Evaluate current practices, organize BP effort
- Push for wider adoption of CLIMSTAN (ClimDB variables) for site met data
    - Organize similar researcher/IM collaborations for other common classes of data
- Tie attribute standardization to LTER synthesis efforts
- Leverage PASTA framework to map site attributes to network attributes as standards emerge
    - Level 0 data = site data
    - Level 1 data = cached data
    - Level 2 data = metadata mapped to LTER attribute standards, …

Discussion

- Other examples, issues
- Prerequisites to starting this process?
- Where should standardization happen (site, PASTA, synthesis projects, end-users)?
- Legacy data/IMS or only new data?
- Who should pay?