The Problem of Semantics in the Metadata Mess

V.M. Megler
David Maier
Portland State University

Figure: CMOP’s “Virtual Columbia River”

With thanks to the scientists at Center for Coastal Margin Observation and Prediction (CMOP). This work is supported by NSF award OCE-0424602.
Agenda

- Our “Big Data” Search Engine
- The Metadata Mess
- Reducing Semantic Diversity
- “Metadata Wrangling”
- Current State
Our “Big Data” Search Engine

• Problem: finding relevant data in a “big data” archive
  ➢ Many datasets, dataset shapes and sizes, physical locations, formats, tools (Megler and Maier, 2011; 2012; 2013)
  ➢ Example information need:
    “observations collected near [lat = 45.5, lon = -124.4] in mid-2010, with temperature between 5-10C”

• Solution: a data search engine that operates over big data archives

Figure: Heterogeneity of Data Formats and Data Access Tools in One Scientific Archive
IR Architecture Adapted to Scientific Data Search

• Approach:
  1. Scan (heterogeneous) data; extract summary features
  2. Search over features, with real-time response
     • Return ranked results, with links to data and tools

Figure: Information Retrieval Architecture, adapted to data search
(from Megler and Maier, 2012)
Ranked Search Over Data: Location, Time, Variables

Figure: “Data Near Here” Search Interface  
(from Megler & Maier, 2011)
Detailed Search Result: Variable Information

- Search result leads to “dataset summary”
- Displays dataset variable information from metadata catalog
- Features produced via one-time scan per dataset

### Data Near Here V0.5: Dataset Details

**Dataset Summary**

<table>
<thead>
<tr>
<th>Field</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agency</td>
<td>Center for Coastal Margin Observation and Prediction</td>
</tr>
<tr>
<td>Description</td>
<td>Foreunner Daily, Foreunner, 2009-05-28</td>
</tr>
<tr>
<td>Type</td>
<td>Cruise</td>
</tr>
<tr>
<td>Data Format</td>
<td>CSV</td>
</tr>
<tr>
<td>Quality</td>
<td>raw_data</td>
</tr>
<tr>
<td>Time: Start</td>
<td>2009-06-28 08:05 PDT</td>
</tr>
<tr>
<td>Time: End</td>
<td>2009-05-28 16:05 PDT</td>
</tr>
<tr>
<td>Depth Min</td>
<td>0m (free surface)</td>
</tr>
<tr>
<td>Depth Max</td>
<td>0m (free surface)</td>
</tr>
<tr>
<td># of Values</td>
<td>2.775</td>
</tr>
<tr>
<td>Data Location</td>
<td>Download</td>
</tr>
<tr>
<td>Last Updated</td>
<td>2011-12-01 08:12 PST</td>
</tr>
</tbody>
</table>

### Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Units</th>
<th>Datatype</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>conductivity</td>
<td>unknown</td>
<td>double precision</td>
<td>unknown</td>
<td>0</td>
<td>0.32</td>
<td>2,774</td>
</tr>
<tr>
<td>salinity</td>
<td>unknown</td>
<td>double precision</td>
<td>unknown</td>
<td>0.06</td>
<td>26.54</td>
<td>2,774</td>
</tr>
<tr>
<td>temperature</td>
<td>c</td>
<td>double precision</td>
<td>c</td>
<td>12.23</td>
<td>18.02</td>
<td>2,774</td>
</tr>
</tbody>
</table>
Motivation for This Work

Emerging problem: Many names for same environmental variable*

- E.g.: temperature, temp, water_temperature
- “Semantic diversity”
- Similar problems in other areas, e.g. variable units

* Hereafter called “variables”
The Metadata Mess

- Working assumption: each named column in a (publicly available) file / dataset represents a valid variable

- Result: Ever increasing number of variables (over 300 at CMOP)

- Problem:
  - Hard for searchers to navigate, locate desired variable
  - Not what the archive wants to expose – “metadata mess”

Figure: Variable List as Exposed in Search Tool
Characterizing the Metadata Mess

- Archive curator’s goal: to present the metadata he wishes he had

- Sources of the mess:
  - Poor, unenforced or multiple naming standards
  - Data from multiple or external sources or systems
  - Changes in systems, standards and personnel over time
  - Many researchers, from different fields
  - Changing research foci

- Can’t we repair the archive?
  - Datasets must be modified or regenerated – not practical
  - May require changing code, systems – expensive, limited payoff
  - Names may be set by vendors or external data providers
  - Time-consuming, error-prone – and problems recur
  - Change is constant
The Metadata Mess (2)

- Alternative approach: compensate for the mess

- How?

  - Reduce semantic diversity
    Perfection not needed

  - Provide transformation layer from “what is” to “what should be”
## Categories of Semantic Diversity

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor variations and misspellings</td>
<td><em>air_temperature, air_temperatrue, airtemp</em></td>
</tr>
<tr>
<td>Synonyms</td>
<td><em>C, degC, Centigrade</em></td>
</tr>
<tr>
<td>Abbreviations</td>
<td><em>MWHLA</em></td>
</tr>
<tr>
<td>Excess variables</td>
<td>Quality assurance variables: <em>qa_level</em></td>
</tr>
<tr>
<td>Ambiguous usages</td>
<td><em>temp: temporary or temperature?</em></td>
</tr>
<tr>
<td>Source-context naming variations</td>
<td><em>temperature may mean air_temperature or water_temperature</em>, depending on source context*</td>
</tr>
<tr>
<td>Concepts at multiple levels of detail</td>
<td><em>Fluorescence, vs. fluores375, fluores400</em></td>
</tr>
</tbody>
</table>
Semantic Diversity: Overall Approach

- **Principles:**
  - No one approach sufficient
  - All approaches must be simple; robust; tolerant of continued growth and ambiguity
  - “Refunds and exchanges available”
    - Provide defaults
    - Improve results via overrides, modifications, adjustments
    - Be non-destructive: re-doable metadata processing

- “Semi-curated” model
  - Curator performs some work for each new type of data indexed
  - Curator can review, adjust and override currently-used defaults and prior decisions
# Reducing Variable-Name Diversity: Possible Approaches

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Desired Result</th>
<th>Possible Technical Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minor variations and misspellings</td>
<td><code>air_temperature, air_temperatrue, airtemp</code></td>
<td>Make them the same</td>
<td>Translate current to desired name</td>
</tr>
<tr>
<td>Synonyms</td>
<td><code>C, degC, Centigrade</code></td>
<td>Make them the same</td>
<td>Translate current to desired name</td>
</tr>
<tr>
<td>Abbreviations</td>
<td><code>MWHLA</code></td>
<td>Use full/canonical variable name</td>
<td>Translate current to desired name</td>
</tr>
<tr>
<td>Excess variables</td>
<td>Quality assurance variables: <code>qa_level</code></td>
<td>Exclude from search</td>
<td>Mark variables</td>
</tr>
<tr>
<td></td>
<td><code>temp: temporary or temperature?</code></td>
<td>Show in detailed dataset views</td>
<td>Exclude from search</td>
</tr>
<tr>
<td>Ambiguous usages</td>
<td><code>temp: temporary or temperature?</code></td>
<td>Identify and expose variables. Allow curator to:</td>
<td>Provide interface to specify options</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• clarify where possible</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• hide variable</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• leave as is</td>
<td></td>
</tr>
<tr>
<td>Source-context naming variations</td>
<td><code>Temperature: air_temperature or water_temperature</code> depending on source context</td>
<td>Specify context of variable</td>
<td>Link to multiple taxonomies</td>
</tr>
<tr>
<td>Concepts at multiple levels of detail</td>
<td><code>Fluorescence, vs. fluores375, fluores400</code></td>
<td>Collapse or expose as needed</td>
<td>Allow variables to be grouped</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Support hierarchical menus</td>
</tr>
</tbody>
</table>
Components of “Metadata Wrangling”

- **Scan archive**
  - Configure scanner: directories, file types, naming conventions
  - Often exists as a translation table
  - “The mess that’s left”

- **Add external metadata**
  - Discover transformations

- **Generate hierarchies**

- **External Metadata**
  - Publish

- **Working Catalog**
  - Perform known transformations
  - Perform discovered transformations

- **Archive Datasets**
  - E.g.: scripts accessing a database

- **External Metadata**
  - Configure: levels, aggregation

- **Metadata Catalog**
Set of composable components
- Compose into a “metadata processing chain”
- Details of process are different for each archive
Current State

- Diversity of variable names is an issue – even within a single archive. Even larger issue when searching over federated archives.

- Metadata wrangling is an ongoing activity.

- We have:
  - Analyzed the problem for our archive (CMOP) and data included from other archives.
  - Suggested possible approaches to address.
  - Experimented with components of the process (scanner; hierarchy generator; scripts to add metadata; discovering & applying transformations).

- Giving a data curator tools to manage what she exposes – to manage her metadata mess – we enable easier use of her data archive.

- By combining this work with our search engine, we allow more effective discovery, access and use of the archive’s contents.
References


Major Curatorial Activities

1. Creating process
2. Running (or rerunning) process
3. Improving process
   E.g., modifying a hierarchy, adding entries to a synonym table, specifying an additional directory to scan
4. Validating process results
   E.g., verifying that all files in a certain directory were indeed of the same type; checking that all harvested variables names occur in the current synonym table as preferred or alternate terms; determining that expected datasets do indeed show up.
Managing “the Mess that’s Left”

- “Discovered transformations” – discovered by reviewing results so far
  - Experimenting with Google Refine*

- Archive curator:
  1. Accesses list of variables (along with sample datasets they appear in)
  2. Reviews list
  3. Generates set of variable-name transformations and rules
  4. Applies rules and checks results for validity
  5. Exports rules and “applies”

- Transformation Engine:
  - Reruns at intervals: as new datasets are scanned
  - Applies rules to existing metadata

- Search engine:
  - Searches over “cleaned” metadata

---

* [http://code.google.com/p/google-refine/](http://code.google.com/p/google-refine/)
Discovering Transformations with Google Refine

Working Catalog

Extract catalog entries to Google Refine

Export JSON Rules

Google Refine Transformations

Run rules against metadata

---

{ "op": "core/mass-edit",
  "description": "Mass edit cells in column field",
  "engineConfig": { "facets": [],
    "mode": "row-based" },
  "columnName": "field",
  "expression": "value",
  "edits": [ {
    "fromBlank": false,
    "fromError": false,
    "from": [ "ATastn" ],
    "to": "sea surface temperature" } ] },