Ontology-based Metadata Portal for Unified Semantics (OlyMPUS)

ESTO Earth Science Technology Forum 2016

June 15, 2016

Jonathan Gleason

jonathan.l.gleason@nasa.gov

Beth Huffer

beth@lingualogica.net

OlyMPUS Team

Principal Investigator

Jonathan Gleason

Co-Investigator / Science Principal Investigator

Elisabeth B. Huffer

Co-Investigator

Pamela E. Mlynczak

Collaborators

Gerald Potter

Mark McInerny

Piyush Mehortra

Pamela Rinsland

Agenda

- Project Overview
- Ontology-based Data Model
- Metadata Provisioning Capability
- Using AIST Managed Cloud Environment Precursor
- Future Work

OlyMPUS Overview

- Ontology-based Metadata Portal for Unified Semantics (OlyMPUS) is a data access, data delivery & metadata provisioning platform

- Goal: To enable researchers to seamlessly find and retrieve variables of interest from multiple disparate datasets and create custom data subsets tailored for their particular research needs

OlyMPUS Overview

Project Objectives:

- Significantly enhance search functionality of existing Ontology-Driven Interactive Search Environment for Earth Science (ODISEES) tool
- Build semi-automated metadata provisioning utility to populate ontology with semantically rich metadata
- Integrate OlyMPUS variable-level search capability with CERES and MERRA subsetting/ordering tools
- Integrate OlyMPUS platform with NEX environment: NEX users can directly pull ordered subsets directly to NEX file system on Pleiades or the NEX sandbox

Data Model

- Using an Ontology-based data model implemented as an RDF triplestore

- The data search tool and the metadata provisioning tool interface with the triplestore back-end that is queried based on user input

- Everything stored as a triple – an assertion of the form <subject> <predicate> <object>

Data Model

Why use an ontology-based data model like this?

- Earth science data sets are diverse, complex and evolving
  - Structure of new data sets often dependent on measurement type
  - As the state-of-the-art advances, we'll likely learn of new relationships among datasets
  - Metadata standards will continue to evolve
- Triplestores provide flexibility for schema changes and new data types
  - Any event or entity can be the subject or object of a triple & any relationship can be a predicate
  - Schema or content updates easy – just add new triples
- Once populated, its often difficult to modify the schema for a relational DB

The Triplestore Advantage

- Provides for comprehensive, structured queries
- Triplestores offer the ability to interpret data and reason over data to discover new facts and relationships not already present in the database
- Schemas are semantic in that they provide formal, machine-readable definitions of the meaning of object types and relationships
- The use of a triplestore supports OlyMPUS end game objective to enable multi-variate analysis
- Supports longer-term goals:
  - Integrate text mining to link documents, especially research articles
  - Create rule-based mappings between data sets and analytical tools and/or data services

UID that identifies a particular variable in a MOPITT data set; this refers to the object to which a set of attribute/value pairs can be ascribed

:MOP03JM-RetrievedCOMixingRatioProfileDay-V006

:variableName	:RetrievedCOMixingRatioProfileDay ;
:parameter	:CarbonMonoxide ;
:mediumType	:Air ;
:quantityType	:VolumeticContent ;
:unitOfMeasure	:PartsPerBillionByVolume ;
:verticalLocationType	:Atmosphere-Profiles .
:profileType	:Atmosphere-PressureLevels ;
:spatialResolution	:GridCell-1x1 ;
:dataSet	:MOP03JM ;
:dataSource	:SatelliteObservation ;
:gridType	:EqualAngleGrid ;
:instrument	:MOPITT-Instr ;
:temporalResolution	:OneMonth ;

Most values are objects themselves that are described by a different set of attribute/value pairs; for example, the data set in which this variable occurs.

Set of attribute/ value pairs that uniquely describe the parent data set (in .ttl format)

```
:MOP03JM
        rdfs:label                        "MOP03JM"^^xsd:string ;
        :archiveCenter                    :NASAASDC ;
        :dataFormat                       :HDFEOS5Format ;
        :doi                                "10.5067/TERRA/MOPITT/DATA301 "^^xsd:string ;
        :grid-HorizontalSpace               :MOPITT1x1Grid ;
        :gridType-HorizontalSpace           :EqualAngleGrid ;
        :instrument                         :MOPITT-Instr ;
        :name                               "MOP03JM"^^xsd:string ;
        :principalInvestigator              :JohnGille ;
        :processingLevel                    :Level3Product ;
        :productionStatus                   :ASDCCurrentDataProduct ;
        :project                            :MOPITTMission ;
        :spatialExtent                      :EarthSurfaceRegion-Global ;
        :temporalCoverage                   :TimeSpan2000-03-03ToPresent ;
        :temporalResolutionActual           :OneMonth ;
```

Values here are also objects that can be described by a set of attribute/value pairs; for example, instrument that collected the data in this data set

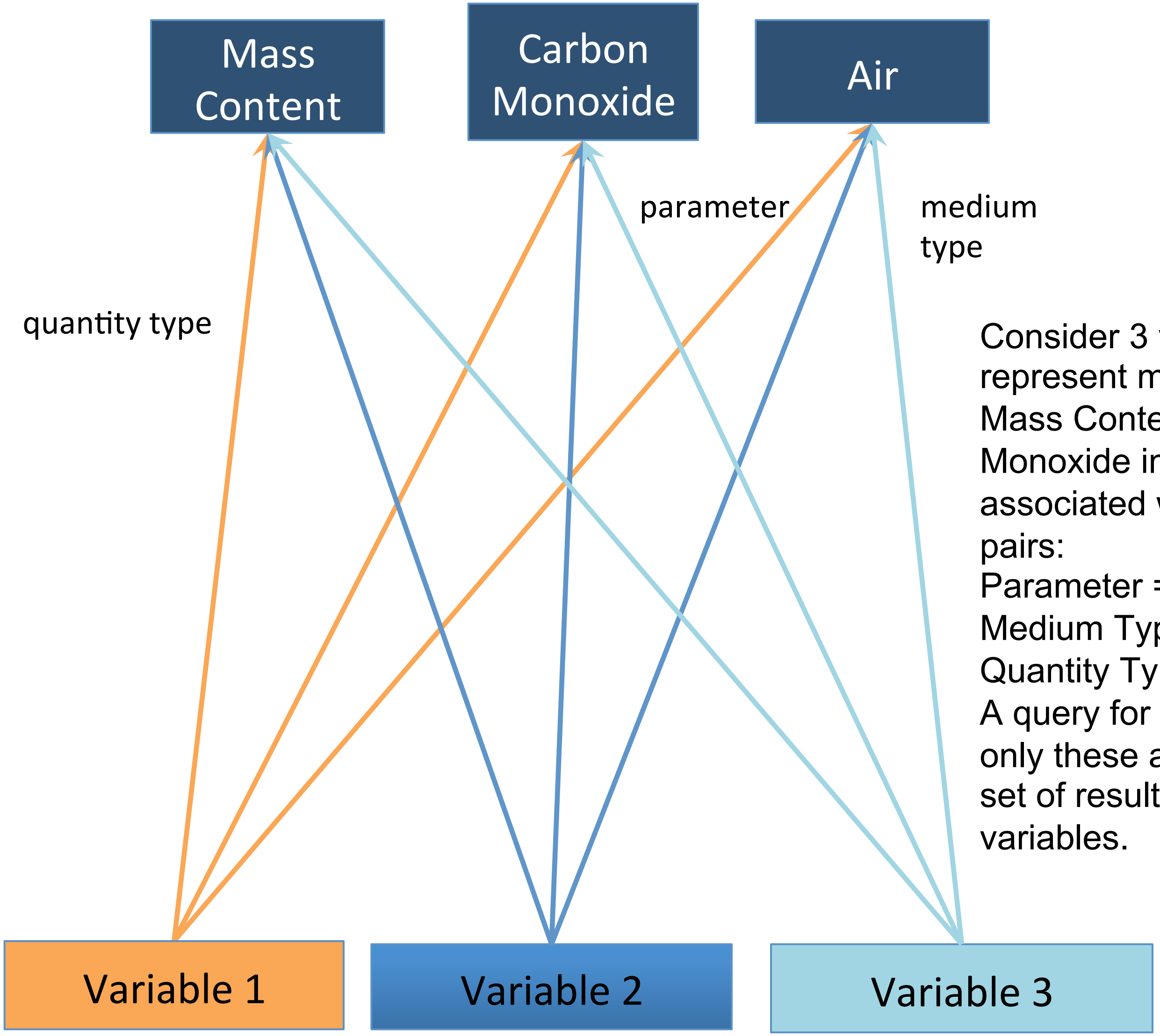
Set of attribute/ value pairs that uniquely describe the instrument (in .ttl format)

```
:MOPITT-Instr
    rdf:type                        :NadirSoundingInstrument ;
    rdfs:label                      "MOPITT (Measurements of Pollution in the
                                   Troposphere)"^^xsd:string ;
    :instrumentSpectralRange       :NIR2Point3Micrometers , :TIR4Point7Micrometers ;
    :instrumentType                :NadirSoundingInstrument ;
    :mission                        :MOPITTMission ;
    :platform                       :TerraSatellite .
```

Set of attribute/ value pairs that uniquely describe the satellite (in .ttl format)

```
:TerraSatellite
    rdf:type             :ResearchSatellite ;
    rdfs:label           "Terra Satellite"^^xsd:string ;
    :altitude            :Altitude705km ;
    :launchDate          "1999-12-18"^^xsd:date ;
    :mission             :TerraMission ;
    :orbitInclination   "98.5"^^xsd:float ;
    :orbitPeriod         :TerraPeriod ;
    :orbitType           :SunsynchronousOrbit ;
    :sensor              :MODISPFM , :MISR-Instr , :CERESFM2XtrkScanner , :MOPITT-
                          Instr , :MODISPFMBand1 , :CERESFM1 , :ASTER-Instr ,
                          :CERESFM2 ;
    :type-Asserted        :ResearchSatellite .
```

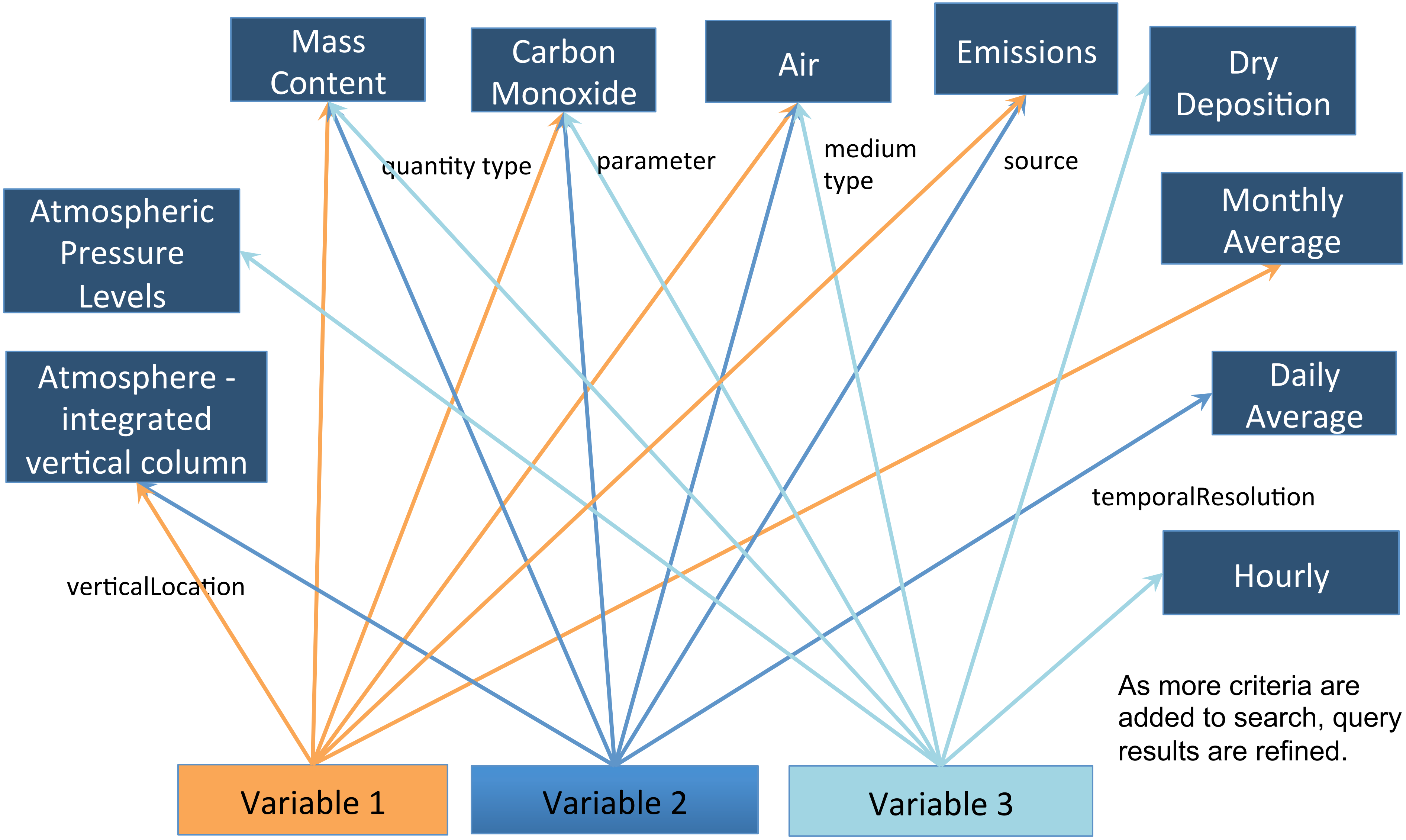
Formal descriptions of variables map similar variables to one another



Consider 3 variables that each represent measurements of the Mass Content of Carbon Monoxide in Air. Each variable is associated with the attribute/value pairs:
Parameter = Carbon Monoxide
Medium Type = Air
Quantity Type = Mass Content
A query for variables with all and only these attributes will return a set of results that includes these 3 variables.

Prospective data users can also see how similar variables differ

Unique characteristic sets per variable with shared characteristics. For example, V1 and V2 are measurements of an integrated vertical column, while V3 is measured at multiple pressure levels.



Mass Content

Carbon Monoxide

Air

Emissions

Dry Deposition

Atmospheric Pressure Levels

Atmosphere - integrated vertical column

Monthly Average

Daily Average

Hourly

quantity type

parameter

medium type

source

temporalResolution

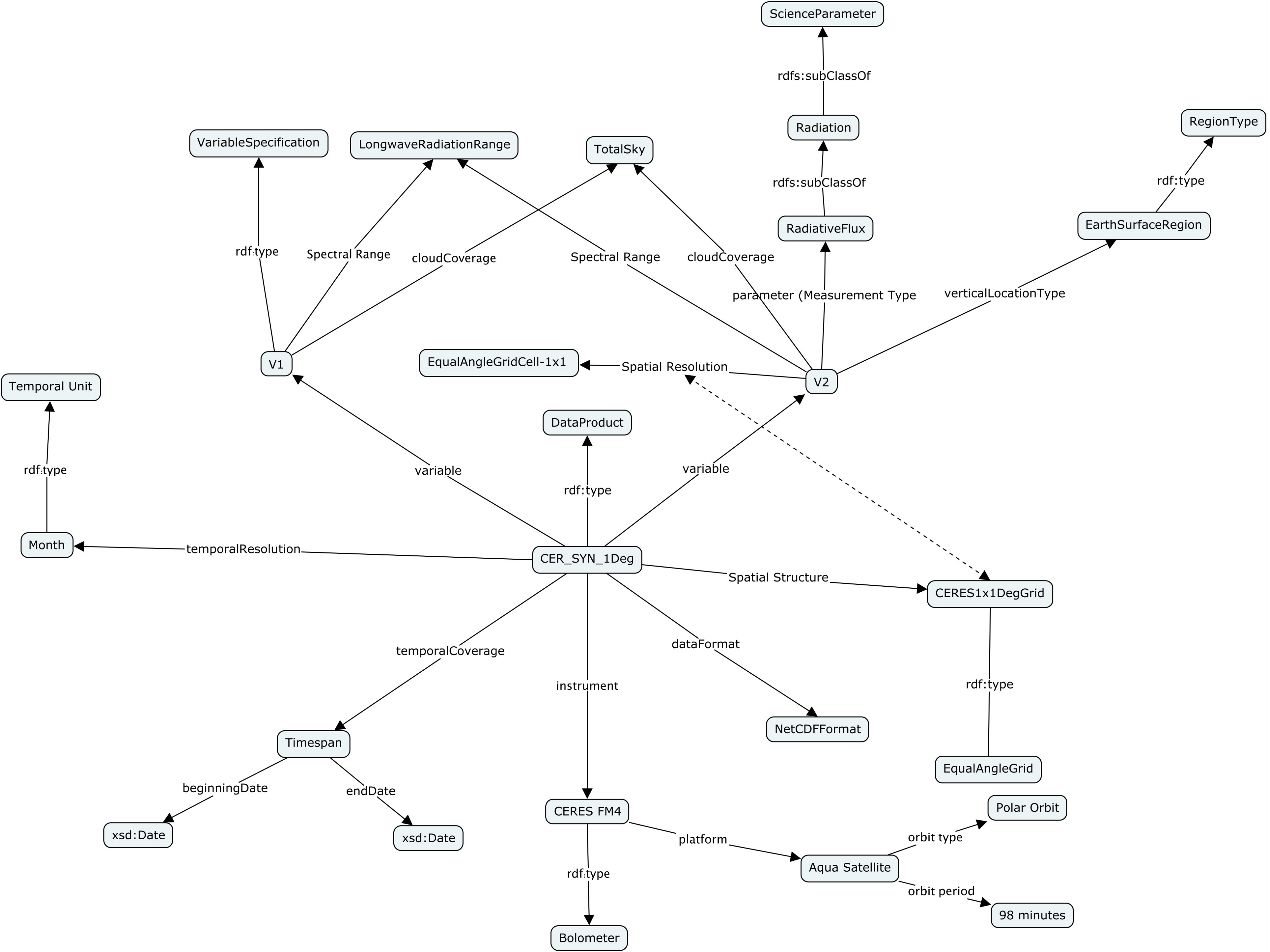
verticalLocation

Variable 1

Variable 2

Variable 3

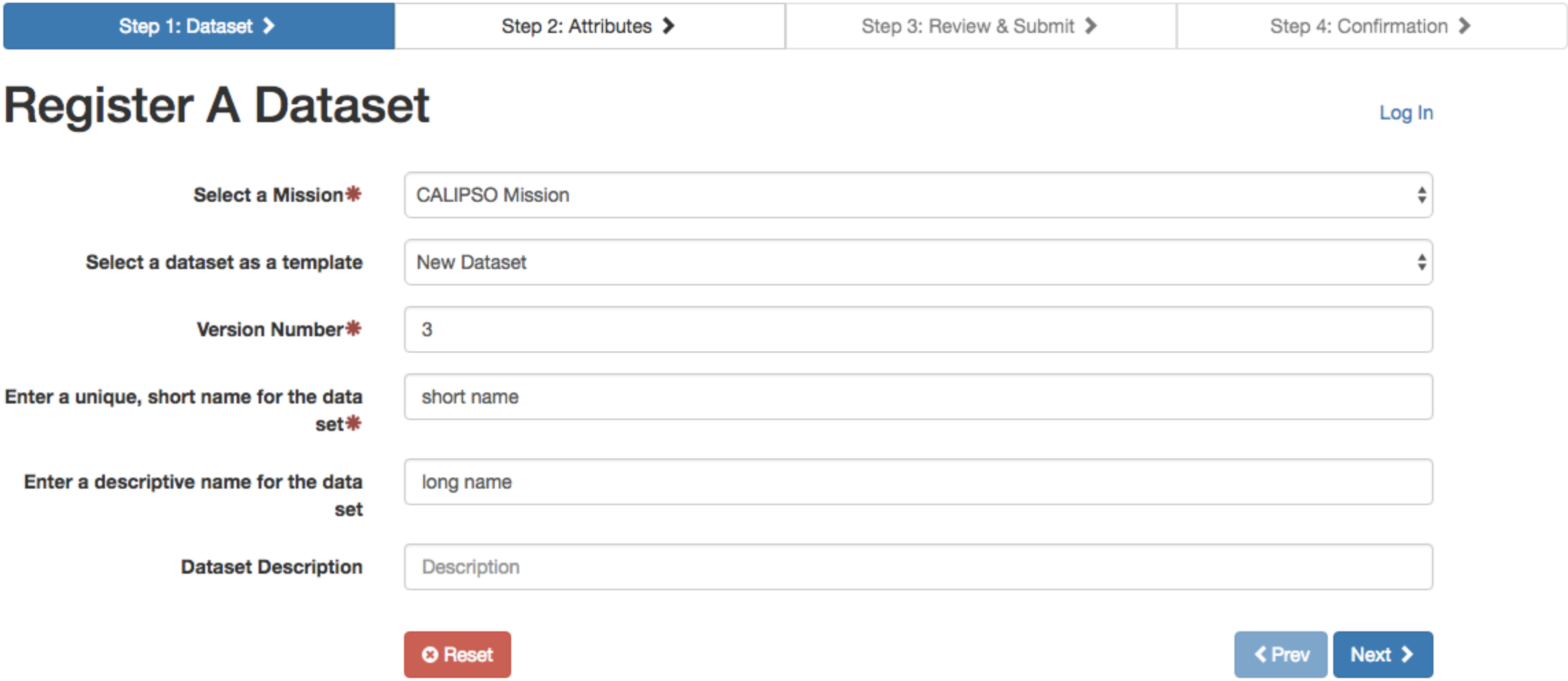
As more criteria are added to search, query results are refined.


ScienceParameter
rdfs:subClassOf
Radiation
rdfs:subClassOf
RadiativeFlux
RegionType
rdf:type
EarthSurfaceRegion
VariableSpecification
LongwaveRadiationRange
TotalSky
rdf:type
Spectral Range
cloudCoverage
Spectral Range
cloudCoverage
parameter (Measurement Type
verticalLocationType
V1
EqualAngleGridCell-1x1
Spatial Resolution
V2
Temporal Unit
DataProduct
rdf:type
variable
rdf:type
variable
Month
temporalResolution
CER_SYN_1Deg
Spatial Structure
CERES1x1DegGrid
temporalCoverage
dataFormat
instrument
rdf:type
NetCDFFormat
Timespan
EqualAngleGrid
beginningDate
endDate
CERES FM4
Polar Orbit
xsd:Date
xsd:Date
platform
orbit type
Aqua Satellite
rdf:type
orbit period
98 minutes
Bolometer

Metadata Provisioning

- Good tools require good metadata
- OlyMPUS prompts users for basic information and uses inference rules with data already in metadata repository to guide users through process to generate new metadata
- Utility reads HDF file and retrieves variable names and structure
- Data registration authentication via EOSDIS User Registration System
- Gaurnteeing accurate and semantically rich metadata requires human input. OlyMPUS combines inference rules and logic to ensure as painless as possible

Metadata Provisioning



Register A Dataset

Select a Mission*: CALIPSO Mission

Select a dataset as a template: New Dataset

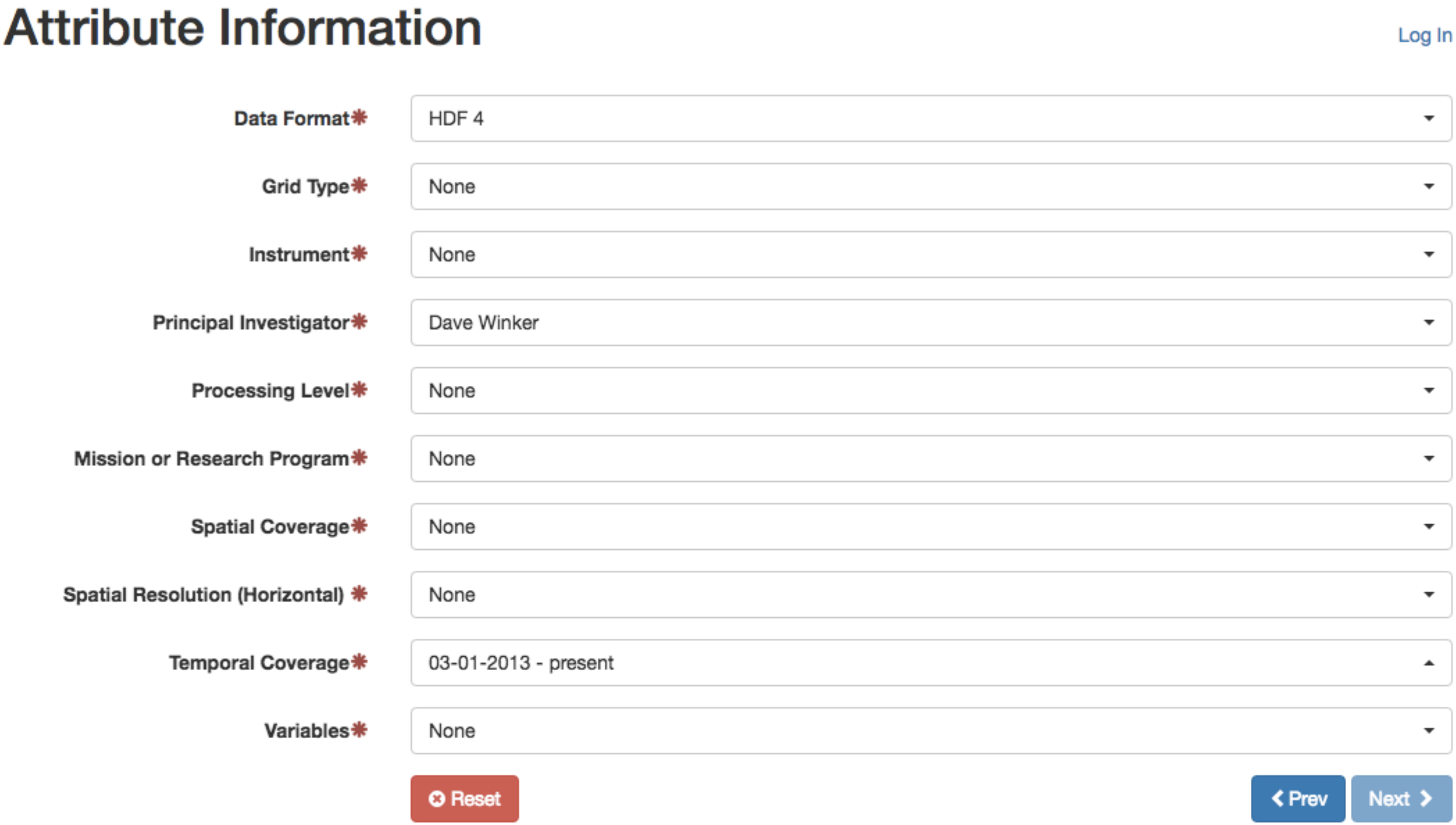
Version Number*: 3

Enter a unique, short name for the data set*: short name

Enter a descriptive name for the data set: long name

Dataset Description: Description

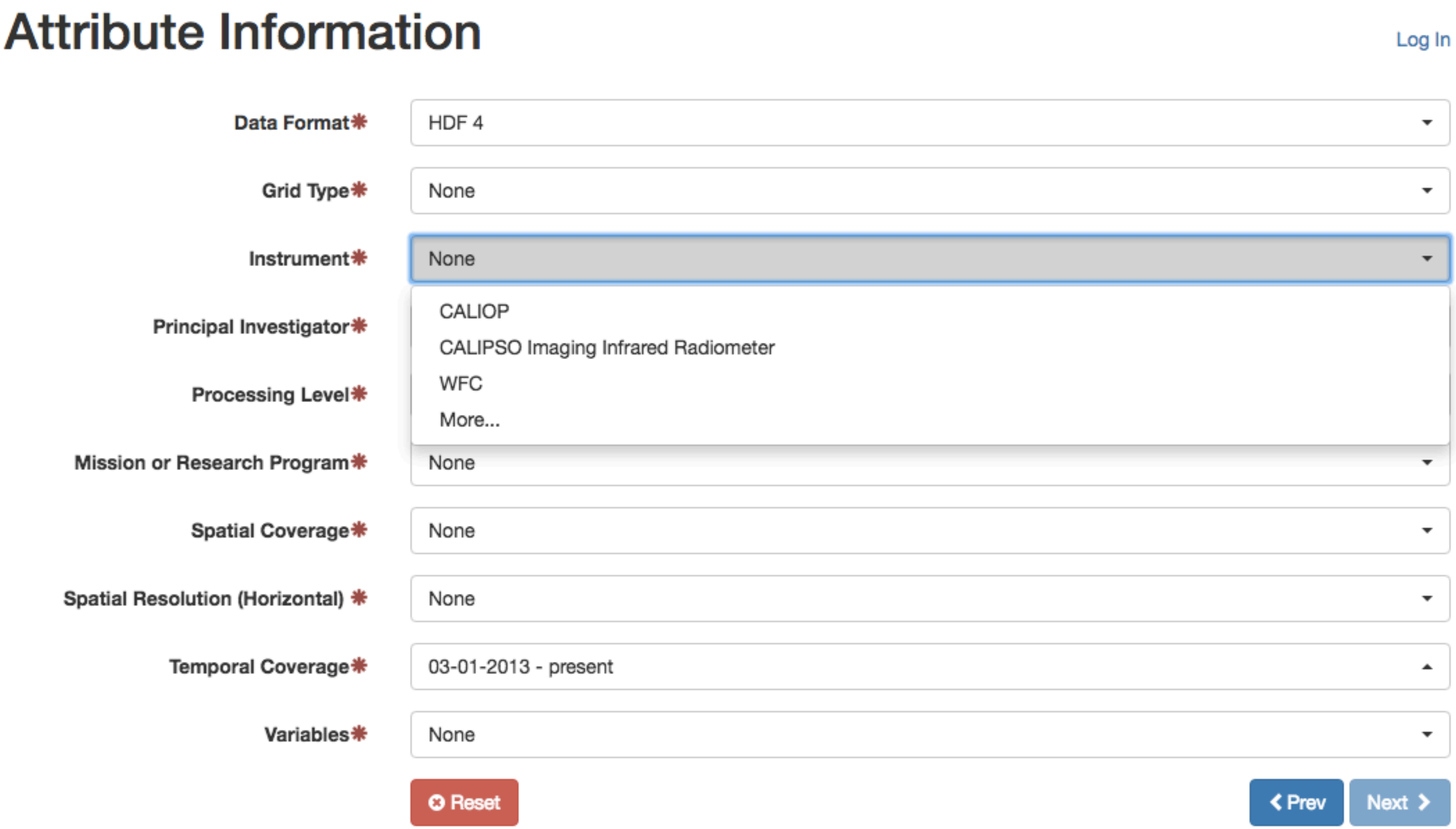
Metadata Provisioning



Attribute Information

Data Format*	HDF 4
Grid Type*	None
Instrument*	None
Principal Investigator*	Dave Winker
Processing Level*	None
Mission or Research Program*	None
Spatial Coverage*	None
Spatial Resolution (Horizontal) *	None
Temporal Coverage*	03-01-2013 - present
Variables*	None

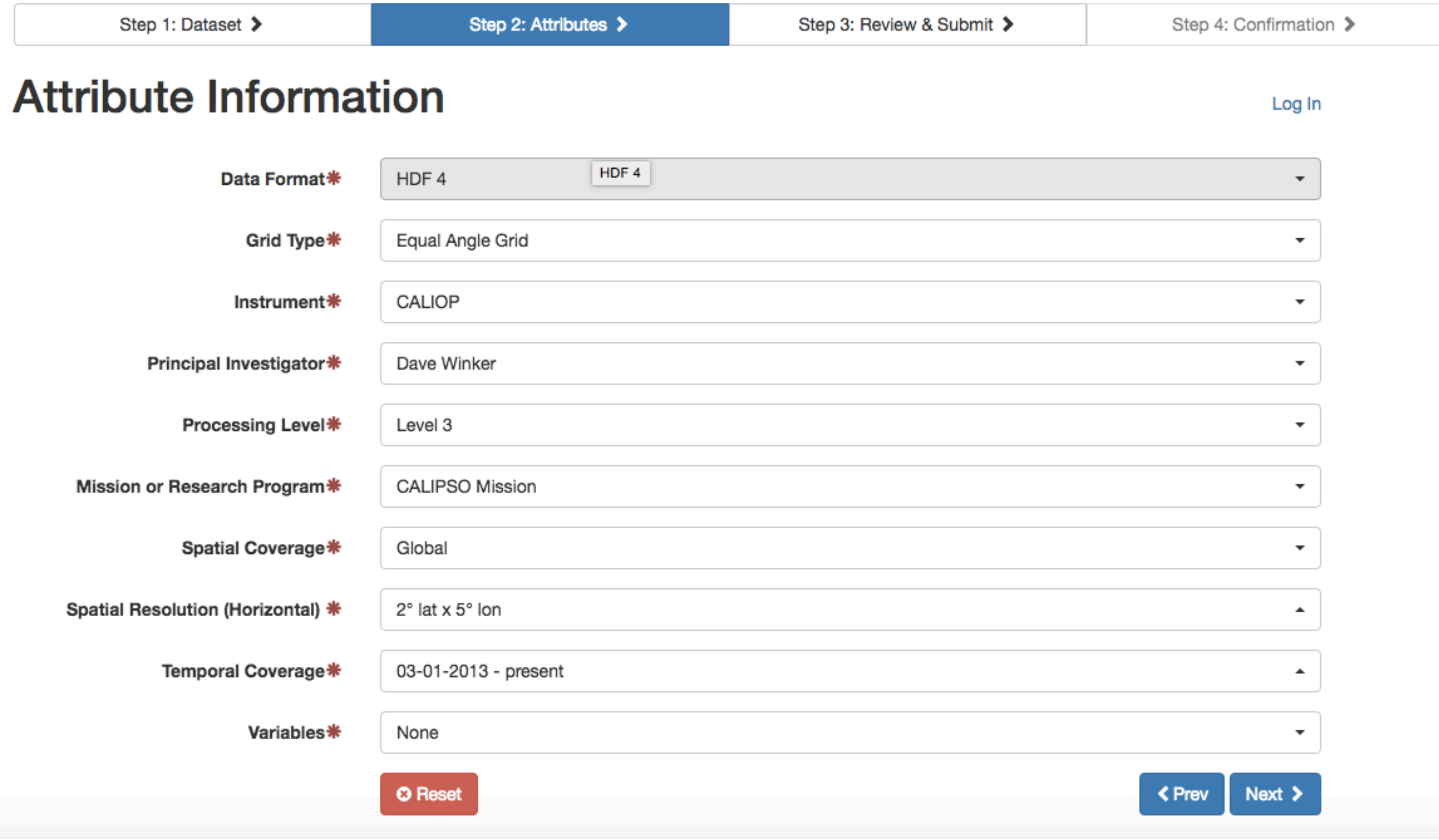
Metadata Provisioning



Attribute Information

Data Format*	HDF 4
Grid Type*	None
Instrument*	None
Principal Investigator*	
Processing Level*	
Mission or Research Program*	None
Spatial Coverage*	None
Spatial Resolution (Horizontal) *	None
Temporal Coverage*	03-01-2013 - present
Variables*	None

Metadata Provisioning



Attribute Information

Data Format*	HDF 4
Grid Type*	Equal Angle Grid
Instrument*	CALIOP
Principal Investigator*	Dave Winker
Processing Level*	Level 3
Mission or Research Program*	CALIPSO Mission
Spatial Coverage*	Global
Spatial Resolution (Horizontal) *	2° lat x 5° lon
Temporal Coverage*	03-01-2013 - present
Variables*	None

Use of AMCE Precursor

- Project development supported by AIST Managed Cloud Environment (AMCE) Benefits / Lessons Learned
- Benefits:
  - Instance management and cost control using DC2
  - On-demand computing more efficient than purchasing and maintaining hardware for smaller projects
- Lessons Learned:
  - AWS Cloud instances help isolate risk from security incidents
  - Incident recovery time, including vulnerability assessment and correction about 3 weeks

Future Work

- Develop an API for micro-service like application
- Develop an API for subsetter interfaces
- Introduce machine-learning techniques to further automate metadata provisioning
- Multi-variate data analysis via this tool

Questions and Comments

General Notes

- Each variable type in a data set is described in detail in a formal language representing the Earth science domain

- Result: User can find all the variables that satisfy some custom-defined set of search criteria and can compare results to see how variables of interest differ

- Data variables are treated as “property bearers” to whivh atomic properties are attributed (Data Model)