



UNAVCO's Participation in NSF EarthCube Project 418: An Initiative in Web-based Geoscience Data Discovery

Using Schema.org/JSON-LD for Describing Datasets

Presented at ESIP 2018 Summer Meeting

Jim Riley, UNAVCO

UNAVCO Data Repository Holdings

UNAVCO archives and provides search and access for data that our community of geodetic scientists uses for quantifying the motions of rock, ice and water that are monitored by a variety of sensor types at or near the Earth's surface.

The data types include:

- GPS/GNSS
- Imaging (point clouds) from SAR and TLS
- Borehole strain and seismic
- Meteorological

GPS/GNSS data sets, TLS data sets, and InSAR products are assigned digital object identifiers.

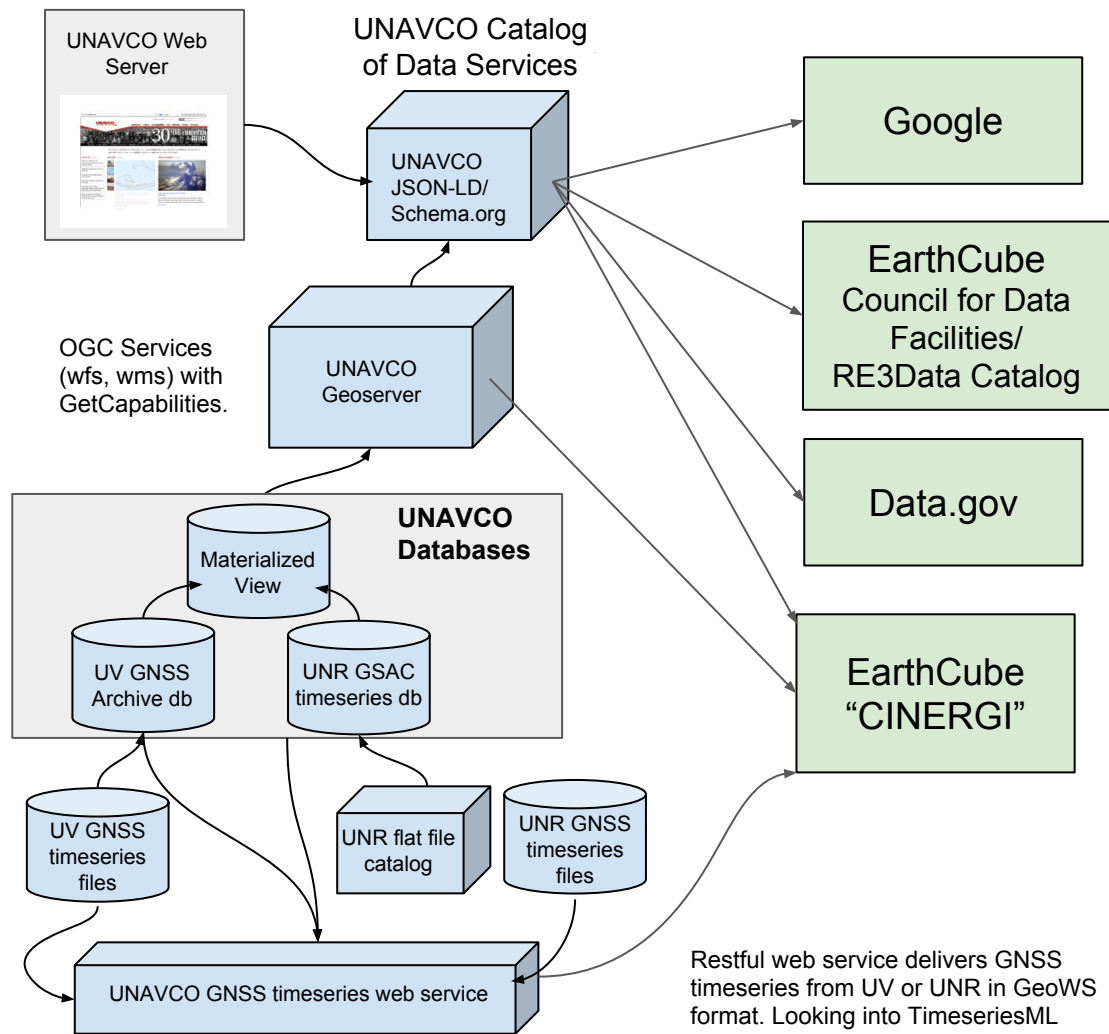
Our Geodetic Data Services group is in the midst of re-architecting our core data systems to better support a variety of geospatial data and real-time data streams.

We are seeking standards for wide interoperability of our metadata within our data center, with partners, and with the world at large.

UNAVCO's Internal GNSS/GPS Dataset Metadata Model

- Used our GNSS/GPS datasets for a pilot project with Project 418 and Google Datasets.
- Already planning to refactor our dataset (DOI) landing pages infrastructure, including a new web service API.
- Utilized existing database tables, in a new materialized view, that supports EZID/DataCite DOI metadata maintenance and our dataset landing pages.
- We have composite and aggregate DOIs for these datasets:
 - Composites have formal two-way hasPart/isPartOf relationship.
 - Aggregates are blind with only one-way hasPart relationship
- For our schema.org structured data we:
 - reviewed P418 full and minimal schema examples
 - reviewed Google Dataset recommendations and examples for schema.org
 - reviewed Schema.org's Dataset schema
 - reviewed existing DataCite JSON-LD output (schema.org) for our datasets via their content resolver
- P418 and Google teams were very helpful with best practices, recommendations, and validation.

UNAVCO's Internal GNSS/GPS Dataset Metadata Model



Primary Services:

- 1) Catalog developed as part of the EarthCube Project 418, headed by Doug Fils (Consortium for Ocean Leadership). Intent is to inform external catalog services, and other users, of UNAVCO's capabilities utilizing Schema.org vocabularies to build structured data (JSON-LD) that is indexable by third parties such as RE3Data and Google.
- 2) Use Open Geospatial Consortium (OGC) compliant Geoserver services to output JSON-LD structured data with general information about data types, holdings, temporal and spatial coverage. Also feed Community Inventory of EarthCube Resources for Geosciences Interoperability (CINERGI) Data Portal through OGC Web Feature Service (WFS). WFS Interface Standard provides an interface allowing requests for geographical features across the web using platform-independent calls. This provides basic individual site metadata and links to web service for GNSS timeseries from UNAVCO or UNR (University of Nevada, Reno). Also provide many of the standard OGC services such a Web Mapping Service (WMS) and Web Mapping Tile Service (WMTS).
- 3) UNAVCO GNSS timeseries web service provides EarthCube's Geoscience Web Services (GeoWS) compliant csv format response. Could be enhanced to provide timeseriesML or netCDF as well. Will need to modify to include a more complete attribution with source data DOI and timeseries DOI.

To Do

- Looking for authoritative ontology to utilize standards where they exist.
- **variableMeasured** and **measurementTechnique** - looking into Geoscience Standard Names (GSN), SWEET, etc. to define and constrain these.
- **keywords** - GCMD?
- Expand schema.org/JSON-LD usage across all of our dataset landing pages.
- Investigating relationship of schema.org to OGC/W3C standards including:
 - Prototyping data sources for Geoserver using materialized db views with basic schema from Schema.org Dataset
 - Learning and exploring:
 - SensorML
 - TimeserieML/Observations and Measurements for time series data
 - Reviewing potential to serialize schema.org/JSON-LD from OGC services
 - Contributing to OGC based GeodesyML (using GML) markup language for exchange of geodetic geospatial data

Usefulness of Schema.org Dataset Model

- Schema.org/Dataset provides a good basic metadata core. Good for high level common characteristics that are shared across thousands of otherwise diverse datasets.
- Still working towards P418's additionally recommended fields, and their full schema example.
- Straightforward to implement by creating a mapping to existing metadata and can provide a basic model for establishing a metadata schema for a new data store.
- Our prior DataCite work mapped well to schema.org/Dataset. Schema.org metadata can be converted into and out of DataCite's Metadata Schema.
- Investigating utility of schema.org vs. OGC/W3C for internal representations of metadata.

Priorities for Improving Schema.org Dataset Model

- Schema.org currently doesn't support grants/awards, only funders, though there is now a pending extension for a Project class with specific types for FundingAgency, FundingScheme, ResearchProject that all have an Organization property called ownershipFundingInfo.
- Validation of schema implementations.