Leveraging VIVO to connect people, instruments, and data semantically in the geosciences: The UNAVCO case study

Gross, M. Benjamin[1], Rowan, Linda R.[1], Mayernik, Matthew[2], Corson-Rikert[3], Jon, Khan, Huda[3]

1) UNAVCO, Boulder, CO 2) National Center for Atmospheric Research, Boulder, CO 3) Cornell University, Ithaca, NY

About UNAVCO

UNAVCO is a non-profit university-governed consortium funded by the National Science Foundation (NSF) and The National Aeronautics and Space Administration (NASA).

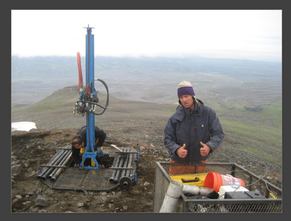
- 106 member universities in the USA
- 94 associate members worldwide

UNAVCO provides data and expertise relating to an array of geophysical methods including GPS, Terrestrial Laser Scanning (TLS), lidar for airborne laser swath mapping, borehole seismometers, strainmeters, and tiltmeters, and interferometric synthetic aperture radar (InSAR).

Researchers from institutions around the world use UNAVCO instruments, data, and expertise to study the earth and how it is changing.

UNAVCO provides services that support research informed by geodesy (the study of Earth's shape, gravity field, and rotation)

Figure 1 (left): Ben Pauk and Kelly Kore conducting repair work on a GPS station in the Cook Inlet near Augustine. UNAVCO maintains an extensive network of continuously operating GPS stations and the associated data across all seven continents. **Figure 2 (below left):** The UNAVCO Polar Team surveys a glacier using a terrestrial laser scanner. Photos from the camera mounted on top of the scanner can be combined with the TLS image. **Figure 3 (below right):** Field engineer Tim Dittmann drilling tiltmeter holes on Unimak Island, AK. UNAVCO has installed a host of borehole instruments as part of the Plate Boundary Observatory.

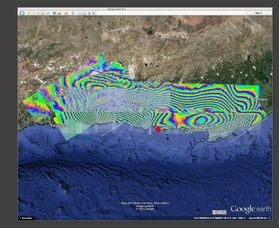


Figures 4 (above left): GPS receivers undergo testing on the roof of the UNAVCO facility. UNAVCO tests both hardware and software and makes recommendations to manufacturers and the research community. **Figure 5 (above right):** The UNAVCO data center hosts community GPS/GNSS data and data products (GPS time series and velocities), high-rate (1-Hz and higher rates) GPS data, SAR data, TLS data, and related metadata for long-term secure storage and retrieval. **Figure 6 (left):** Dave Mencin sits with land owners in central Asia. UNAVCO provides expertise in permitting and land owner relations, as well as education and outreach.

UNAVCO Services

Research community reporting

Research community produces work supported by UNAVCO



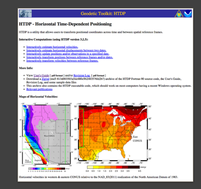


Figure 7 (top left): Examples of academic articles published by a UNAVCO community member at Cornell University. **Figure 8 (top right):** Google Earth KMZ file released by members of the Satellite Geodesy research group at Scripps Institution of Oceanography et al. **Figure 9 (bottom right):** Short-course participants learn how to use UNAVCO data to model volcano deformation. **Figure 10 (bottom left):** The developers of HTDP, a program that is used to transform coordinates across time and spatial reference frames, used tectonic plate velocities estimated by UNAVCO.

About EarthCollab

Enabling Scientific Collaboration and Discovery through Semantic Connections, or EarthCollab, is part of the EarthCube Program at the National Science Foundation. EarthCollab includes two use cases: a VIVO implementation at UNAVCO and another at NCAR's Earth Observing Laboratory (EOL). Cornell, where VIVO was originally developed, is also part of the collaborative project.

NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

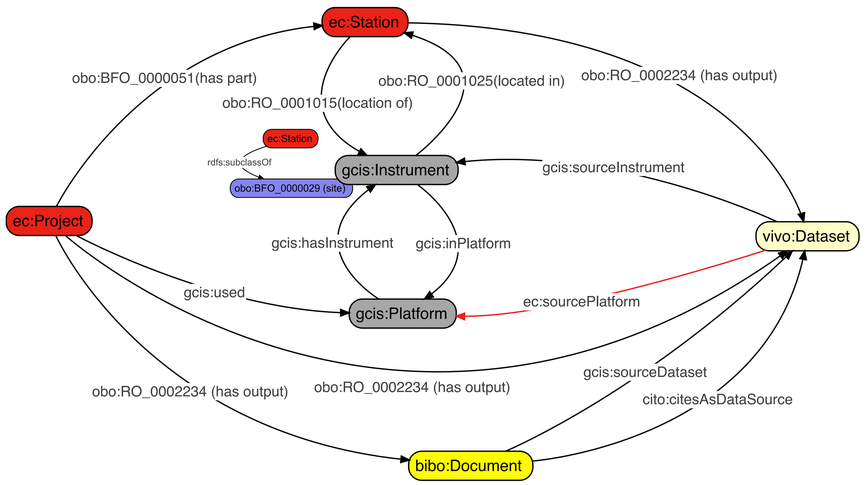
UNAVCO

Cornell University

Data Structures

VIVO has been primarily implemented by research universities, especially in the life sciences. This is reflected in the VIVO-ISF ontology. To capture concepts relevant to UNAVCO, we must extend the ontology. We will use concepts from established ontologies whenever possible.

- VIVO-ISF: people, organizations, publications. VIVO-ISF is already a compilation of ontology terms, including concepts from the OBO foundry (Open Biological and Biomedical Ontologies), CITO (Citation Typing Ontology), and BIBO (Bibliographic Ontology).
- DCAT: datasets and catalogs. DCAT (the Data Catalog Ontology) was designed to facilitate interoperability between data catalogs published on the web. DCAT contains concepts not captured in VIVO's treatment of datasets as a type of publication, such as landing pages and distribution information.
- GCIS: scientific instruments, platforms, and projects. GCIS (Global Change Information System) includes concepts such as sensors, scientific models, and platforms.
- EC: local EarthCollab ontology. UNAVCO operates hundreds of continuously operating GPS stations. Station is a concept not defined in any of the above ontologies and will be defined in EarthCollab's local ontology.



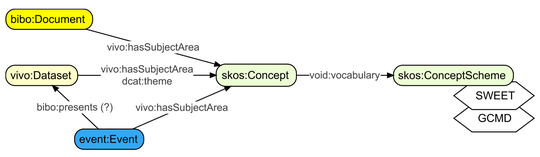
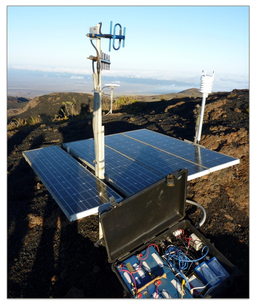


Figure 11 (left): Visual representation of a relationship not fully supported in the VIVO-ISF ontology. A project will have many stations which may host multiple instruments (e.g. GPS and co-located meteorological equipment), which are attached to one or more platforms. See Figure 14.

Figure 12 (above): Datasets can be linked to a major geophysical event, such as an earthquake. The products shown, and others such as software and models, will be tied to skos:Concepts, which will be pulled from controlled vocabularies whenever possible.

Dataset DOIs

Figure 13 (below): We have implemented a Digital Object Identifier (DOI) assignment mechanism, currently for GPS/GNSS data and select inSAR data, and will be expanding their use to other data types. DOIs provide persistent unique identifiers and can ease cross-linking efforts.

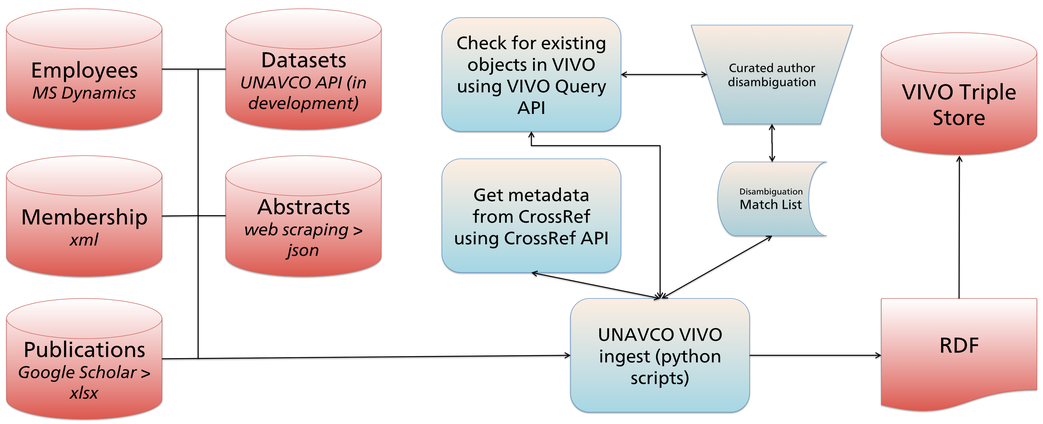


GPS Data Set Citation Summary

Identifier:	10.7283/T57072H2
Title:	Plutons GPS Network
Subtitle:	LLST-Lastarria Volcano P.S.
Authors:	Matt Pritchard
Published:	2011
Publisher:	UNAVCO, Inc.
Description:	Permanent Station
Station Status:	Active
Date Range:	2010-11-06 through 2014-03-23
Citation:	Pritchard, Matt, 2011, Plutons GPS Network: LLST-Lastarria Volcano P.S., UNAVCO, GPS Data Set, doi:10.7283/T57072H2
Related Publications:	[none]
Data Availability:	Anonymous access
Data Access:	http://www.unavco.org/data/gps-gnss/data-access-methods/dai2/dai2.html#/Char=LLST;access=Station;#CharMod=Equals;sampleRate=normal
See Also:	UNAVCO Data Policy UNAVCO Attribution Policy

Figure 14 (left): A completed GPS site on the caldera rim of the Sierra Negra volcano, Galapagos Islands. The station includes an array of equipment types, including the GPS antenna and receiver, solar power source, radio transmitter, and meteorological equipment.

UNAVCO Data Ingest Process



Data Ingest Details

UNAVCO's current data ingest process is illustrated by the flowchart on the left (Figure 15). A major challenge has been collecting data for VIVO from a number of separate databases maintained by separate people.

Ingest process highlights:

- Python code uses rdflib library to allow output of multiple RDF file types.
- Using the VIVO Query API, checks VIVO database for duplicate objects. Unique identifiers, such as DOIs, are used when possible and string matching used otherwise. This module is also used to check for collisions when generating random URIs.
- Includes a rough implementation of a disambiguation process using fuzzy string matching and manual curation. Name variations that are confirmed are stored in a 'Match List' file, which is checked by the script before requesting curator input.
- Currently, triples are loaded into VIVO using the Add/Remove RDF tool. As development continues and we begin to automate ingest, triples will be loaded using the VIVO Update API.

Figure 16 (right): Example of ingest script requesting curator input to match an author's name to names already in the VIVO database.

```
 num  first   middle  last   uri        type   score
 ---  ------  ------  -----  ---------  -----  -----
   0  Ryan S.         Cross  per274395  vcard  87
   1  Paul            Cross  per980517  vcard  70
   2  P               Cross  per118559  vcard  -
   3  P.              Cross  per678453  vcard  -
   4  PA              Cross  per621912  vcard  -

Author Ryan Cross may already exist in the database. Please choose a number
```

Progress

Figure 17 (right): UNAVCO has made significant progress identifying data sources and ingesting the data into VIVO. Additionally, we have begun requesting employees and select community members sign up for ORCID, a service that provides a unique identifier, publications search capabilities, and a public API.

Initial Data Ingest Progress	
People	527
Publications	4062
Datasets	3230
Abstracts	437
Organizations	236

Future Work

- Begin cross-linking VIVO instances across institutions.
- Enhance geospatial capabilities of VIVO by extending ontology and application.
- Automate ingest process, including ingest from ORCID.
- Add GPS station data and link to datasets already in VIVO.
- Conduct task-centered usability testing to determine how the VIVO application can be further tailored to a geoscience-centered use case.
- Explore integration with other EarthCube web projects.

This material is based upon work supported by the National Science Foundation under Grant Nos. ICER-1440213, ICER-1440293, and ICER-1440181.

Photos are courtesy of UNAVCO staff members Beth Bartel, Barrett Friesen, Brendan Hodge, David Mencin, and Joe Pettit.