Guiding development of a semantic web app: End-user engagement in the EarthCollab project

Gross, M. Benjamin[1], Johns, Erica M.[2], Rowan, Linda R.[1], Mayernik, Matthew[3], Khan, Huda[2], Daniels, Michael D.[3], Krafft, Dean B.[2]
1) UNAVCO, Boulder, CO 2) Cornell University, Ithaca, NY 3) National Center for Atmospheric Research, Boulder, CO

Shaping the project

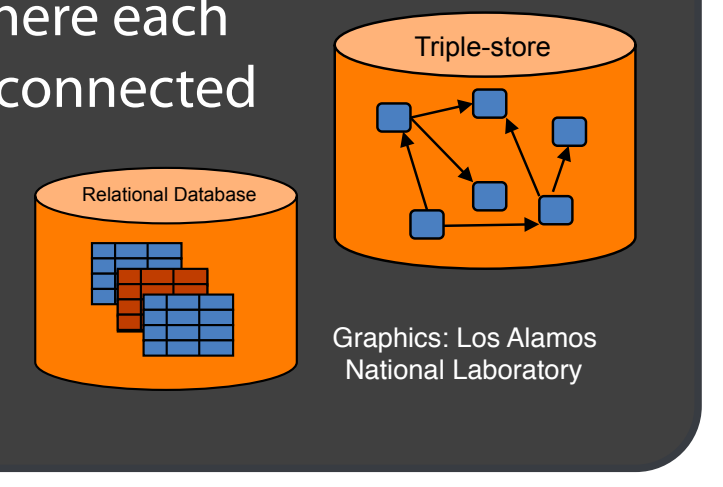
Enabling Scientific Collaboration and Discovery through Semantic Connections, or EarthCollab, is part of the EarthCube Program at the National Science Foundation. EarthCollab has proposed extending an existing open-source semantic web application, VIVO, to highlight connections between people, datasets, grants, and research output.

The project includes two use cases: a geodesy-focused 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.

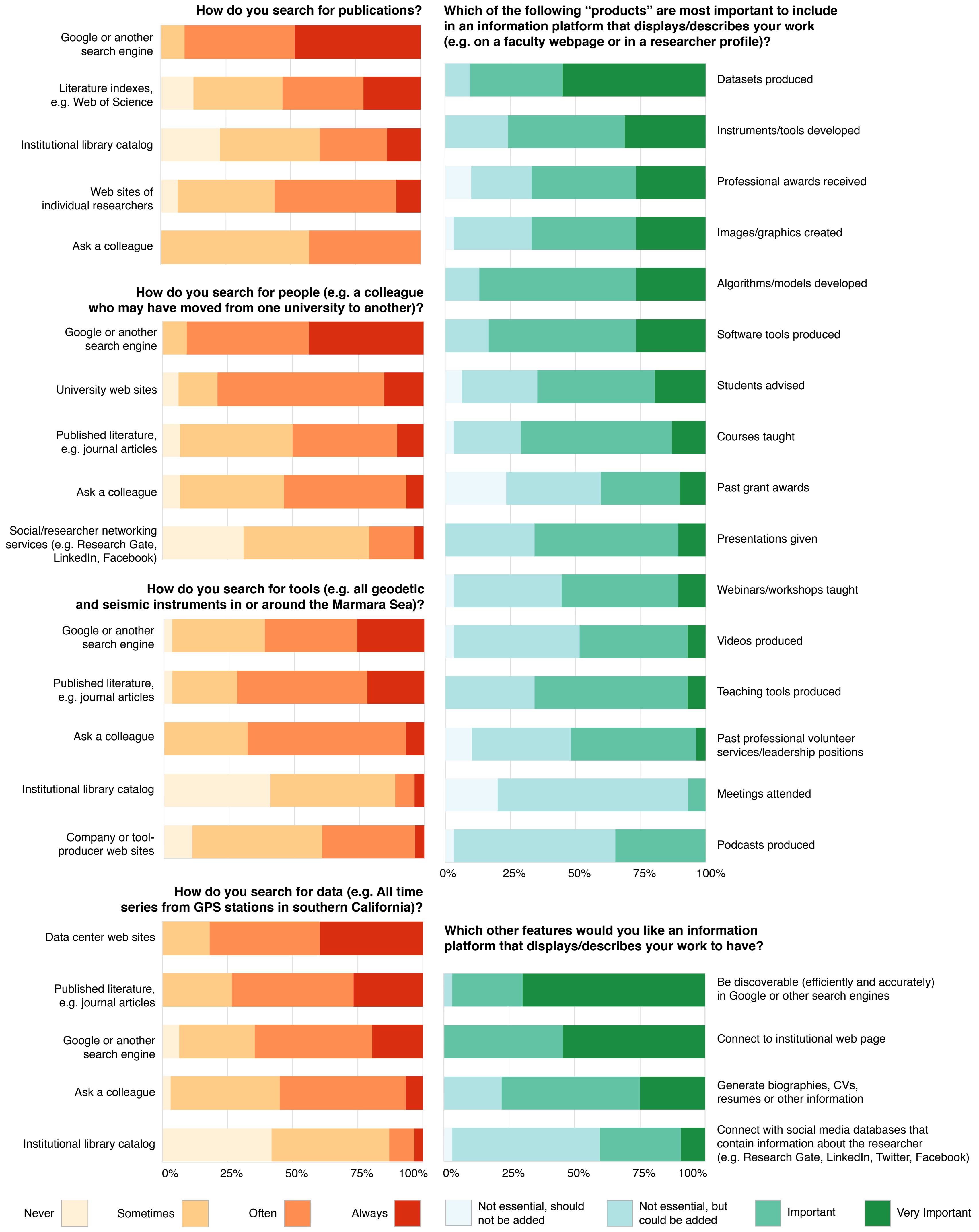
EarthCollab held a workshop at the American Geophysical Union Fall Meeting in December 2014 to identify community needs and how EarthCollab might address them. Workshop participants completed a survey on how they find and share research. The survey, which was completed by 34 researchers including the workshop participants, is summarized below (**Figure 1**).

Why semantic?

Semantic technologies use controlled vocabularies and common formats to store machine-readable data that can be easily reused across applications. Semantic applications store information in triple-store databases. A traditional relational database holds information in columns and rows. A triple store can be thought of as a web, where each piece of information is connected to another according to controlled vocabularies.



Graphics: Los Alamos National Laboratory



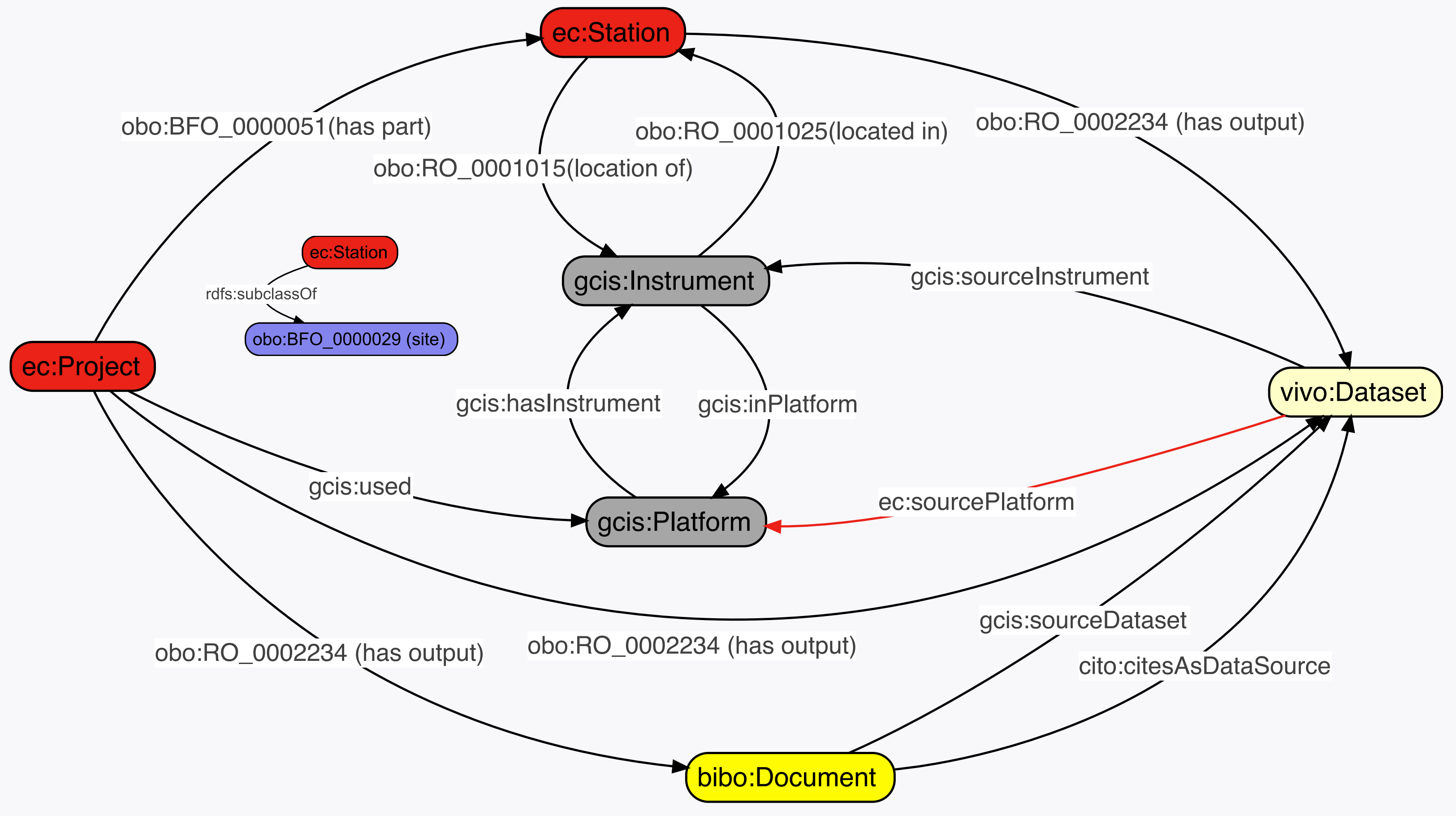
Defining Semantic Connections

Triples are the basic unit of a semantic database. A triple consists of a subject, predicate (verb), and object, each of which are usually described by a uniform resource identifier (URI). For example:

```
<http://connect.unavco.org/individual/org253530> <http://www.w3.org/2000/01/rdf-schema#label>  "UNAVCO" .
```

Subject	Predicate	Object

In plain english, this translates to *organization 'org253530' is named UNAVCO*. The subject of one triple will be the object of another triple, creating a network of linked data.



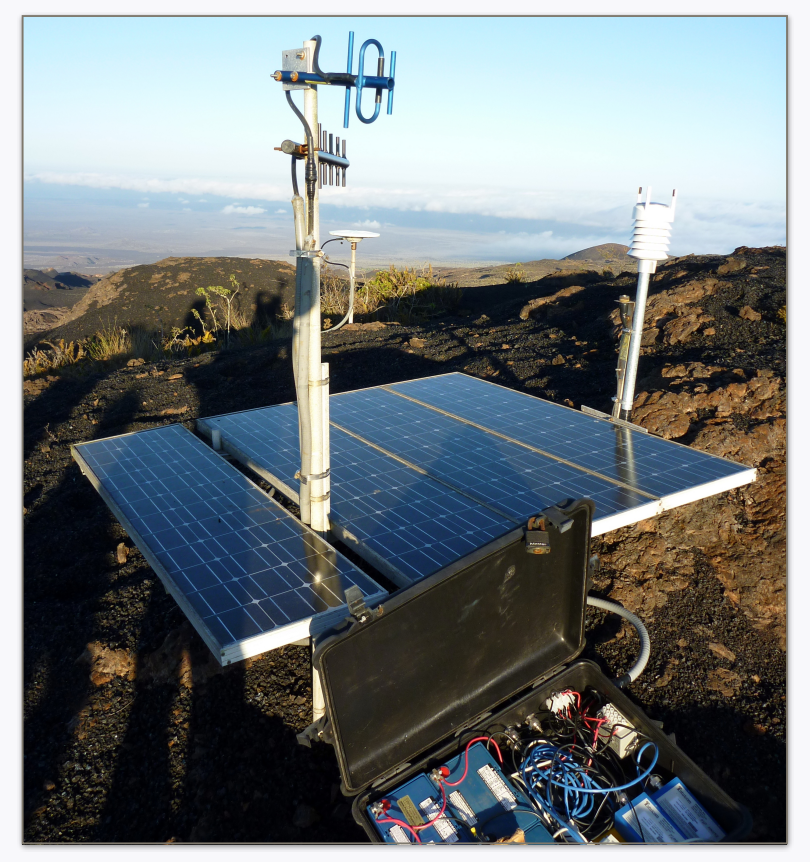


Figure 3 (above): A completed GPS site on the caldera rim of the Sierra Negra volcano, Galapagos Islands. The station includes an array of equipment types: GPS antenna and receiver, solar power source, batteries, radio transmitter, and meteorological equipment.

Figure 2 (above): The VIVO application comes pre-packaged with an ontology (semantic vocabulary) that doesn't cover many geodetic concepts. Figure 2 illustrates 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 meteorological equipment), which are attached to one or more platforms (see Figure 3 for an example).

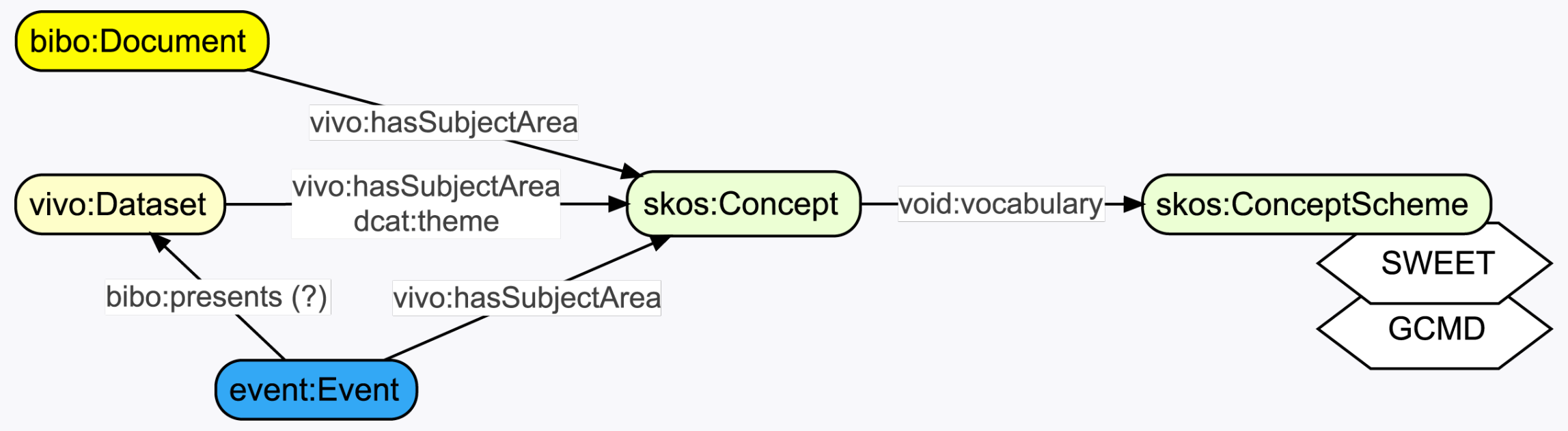


Figure 4 (above): Datasets can be linked to a major geophysical event, such as an earthquake. The products shown, and others such as software and models, are tied to skos:Concepts, which are pulled from controlled vocabularies whenever possible. Geodesy-specific concepts were added to the application to capture research areas and expertise.

Using Unique Identifiers

Unique identifiers make it much easier to connect semantic data. Most peer-reviewed science publications now have Digital Object Identifiers (DOIs) to track and connect publications. Similarly, UNAVCO has recently begun assigning DOIs to datasets through the EZID service. GPS/GNSS and select inSAR datasets are currently being assigned DOIs.

Name ambiguity presents a challenge when populating a semantic database, partly because unique identifiers are not commonly implemented for people. Publication records often only include partial names, making the process of connecting authors with non-unique last names to their unique record in the VIVO database difficult. A publicly available, unique identifier for a person solves this problem, much like a DOI does for an information resource.

ORCID is an open-source, non-profit effort to provide persistent identifiers for people. ORCID has been adopted by publishers such as AGU and AAAS as an optional part of the submission and publication process and is increasingly becoming an essential part of a researcher's academic identity. UNAVCO is leveraging ORCID in the disambiguation process and to facilitate updates to staff and member records using the ORCID public API.

Building out the application

The VIVO application was customized to better capture the needs of the geodesy community. The customizations implemented so far include ontology extensions and mapping capabilities.

For the initial rollout, the database was populated from a diverse array of sources, summarized below (Figure 5). As the application matures, relevant data will be added automatically using APIs from the NSF, ORCID, and CrossRef.

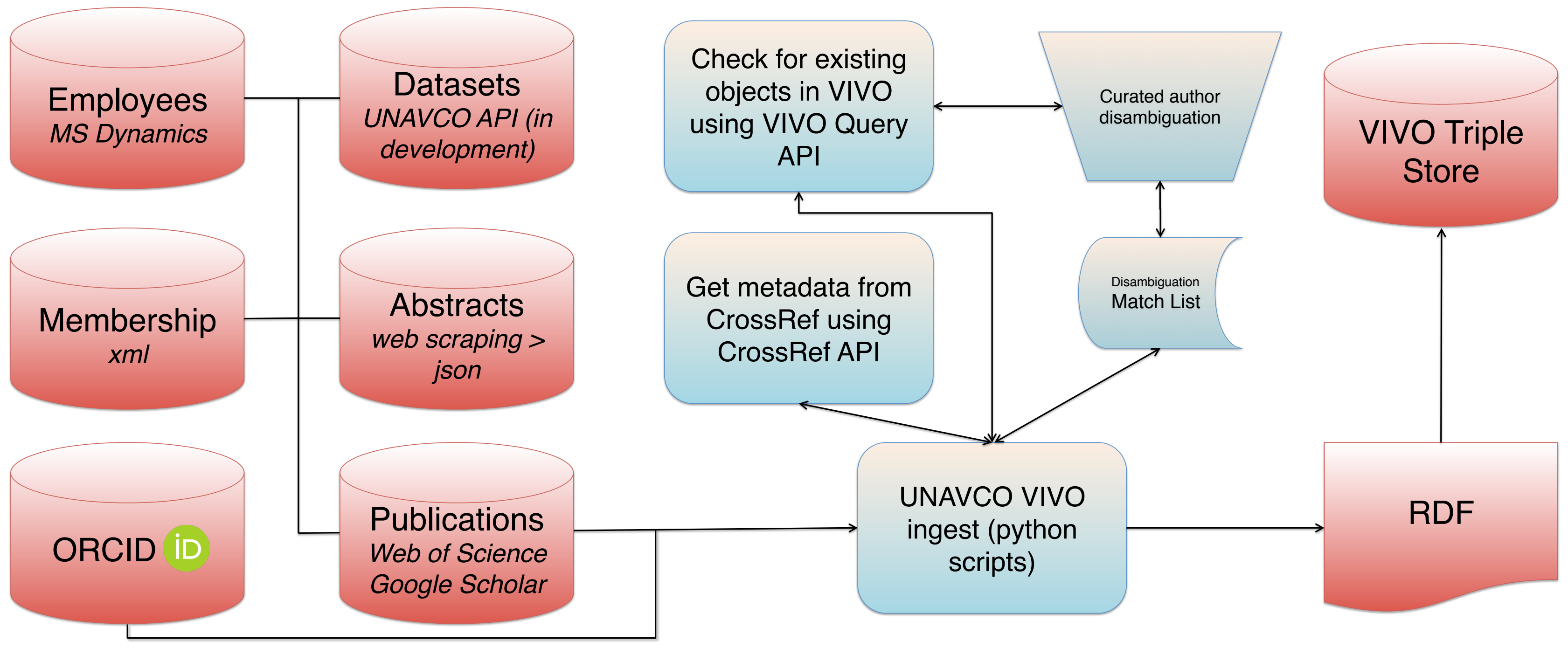

Employees
MS Dynamics
Datasets
UNAVCO API (in development)
Check for existing objects in VIVO using VIVO Query API
Curated author disambiguation
VIVO Triple Store
Membership
xml
Abstracts
web scraping > json
Get metadata from CrossRef using CrossRef API
Disambiguation Match List
ORCID iD
Publications
Web of Science
Google Scholar
UNAVCO VIVO ingest (python scripts)
RDF

Figure 5: Summary of preliminary data ingest workflow. All data must be mapped to an ontology prior to being loaded into the application. More data will be ingested from ORCID as integration with publishers and metadata repositories develops, reducing the need for the intermediate data curation steps indicated in blue.

Connect UNAVCO

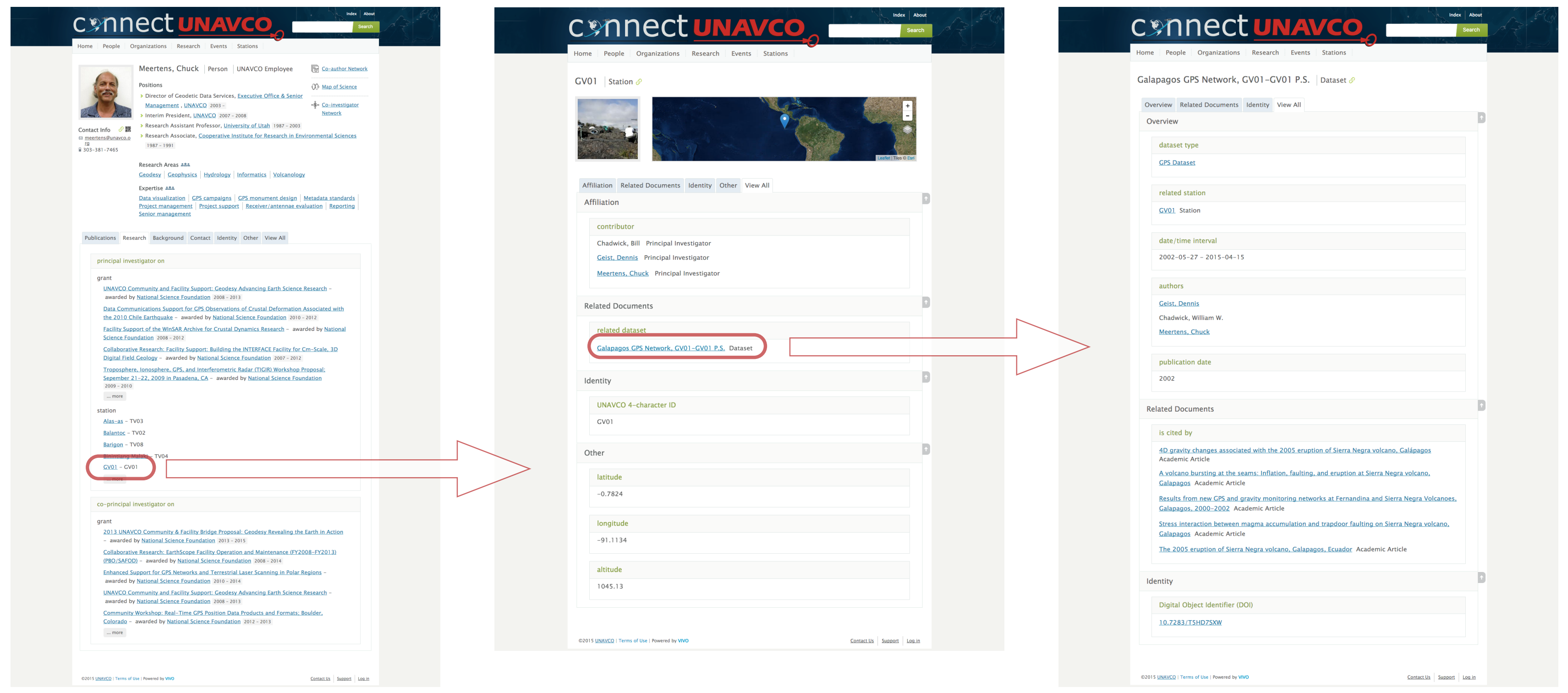


Figure 6: Screenshots of the customized VIVO application, branded as Connect UNAVCO. In this example, UNAVCO Director of Geodetic Data Services Chuck Meertens is a principal investigator on a permanent GPS station named GV01. The station has a related dataset titled Galapagos GPS Network, GV01 - GV01 P.S. The dataset page includes a list of related publications, dataset authors, and date information, as well as the dataset's DOI. All the information displayed is semantically linked together and is available in machine-readable RDF format.

http://earthcube.org/group/earthcollab

GitHub http://git.io/vG9AJ

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.
- Explore integration with other EarthCube web projects.
- Continue customizations based on feedback from usability testing.

Check it out at http://connect.unavco.org

This material is based upon work supported by the National Science Foundation under Grant Nos. ICER-1440213, ICER-1440293, and ICER-1440181. Additional travel support provided by the EarthCube Early Career Travel Grant program.