UNAVCO Data and CyberInfrastructure Efforts on Behalf of the Geodesy Community

Fran Boler, Chuck Meertens and Adrian Borsa (UNAVCO, Boulder, Colorado)

With the viewpoint that the EarthCube process can benefit from a survey of existing cyberinfrastructure efforts, this white paper focuses on the experiences and current priorities at UNAVCO in its ongoing infrastructure development on behalf of the geodesy community. Governance, Data Life Cycle Support, Interoperability Priorities, and Integration Priorities are discussed.

Like CUAHSI, IRIS, and Unidata, UNAVCO is an example of a community-led organization that builds infrastructure to meet the needs of a cohesive disciplinary geoscience domain, in this case geodesy. UNAVCO has benefitted from relatively stable funding through core sponsorship provided by NSF and NASA. While UNAVCO started as a consortium to advance high precision Global Positioning System based geodesy, its scope has expanded significantly to advance applications such as deformation of ice, the Earth's response to ground water, sea level, and other aspects of the hydrosphere, and imaging the structure of the atmosphere. At the same time, the toolbox available to the science community and supported by UNAVCO has expanded to include many new geodetic tools: advancing GPS towards mm-level global GPS geodesy and to streaming high rate observations; borehole strain meters and seismometers, expanded geodetic imaging using LiDAR (Light Detection and Ranging) for Airborne Laser Swath Mapping, InSAR (Interferometric Synthetic Aperture Radar), and Terrestrial Laser Scanning. Meanwhile, GPS is finding applications in a frequency range that used to be the sole provenance of seismology, as GPS moves from one solution per day to one solution per second, with high precision.

UNAVCO provides science support through community coordination, field engineering, data services, technology innovation, and instrument testing, acquisition, and deployment. Further, it supports state-of-the-art global geodetic infrastructure that is developed and operated through international collaborations. EarthScope, a set of integrated geophysical observatories, supports investigation of the entire temporal spectrum of Earth deformation processes. In particular, the Plate Boundary Observatory (PBO) provides unprecedented geodetic imaging of plate boundary deformation. UNAVCO is committed to enabling efficient testing, adoption, and implementation of rapidly evolving geodetic technologies needed to support cutting edge geodynamics research.

UNAVCO's Data Center, with 28 terabytes of GNSS data, 20 terabytes of InSAR data, 9 terabytes of LIDAR data and several terabytes of borehole geophysical and other data, is a strong resource for the geodesy community and for related science disciplines. UNAVCO's Education and Outreach Division works to bring exciting science results from geodesy to the public and to foster data-based learning tools at all levels of education. The UNAVCO Data Center continuously strives to develop and enhance the data and cyberinfrastructure tools available to the geodesy community and education and outreach community.

I. Community Drivers for Data and Cyberinfrastructure Efforts at UNAVCO

NSF's EarthCube process has recognized the importance of the governance model to be applied for overall guidance and buy-in for the EarthCube effort. UNAVCO's community governance structure includes the Board of Directors plus a number of committees and advisory groups whose advisory role touches on data or cyberinfrastructure concerns. The UNAVCO community has endorsed an open data policy for the high precision Global Navigation Satellite System data that forms the core of UNAVCO's data collections. To the extent possible with data collected in related domains, such as InSAR data, where the source is generally a non-US-based space agency such as ESA, JAXA, and others, with guidance from the WInSAR Executive Committee (WInSAR is the Western North America InSAR Consortium, which operates as a self-organizing consortium reporting the UNAVCO Board of Directors), UNAVCO makes data as widely available as possible within the constraints of the data policies at the source agency. Over the past 18 months, the UNAVCO Board has recognized that in concert with an open data policy, there is a need for greater recognition for the significant planning and execution efforts expended by individual investigators in collection of data sets; this recognition has led to increased focus on establishing and implementing data citation support through digital object identifier technology and related data citation guidance to be provided by the UNAVCO Data Center.

UNAVCO CyberInfrastructure: Community Support
• Community governance • Community-developed Strategic Plan • Support for Open Data Policy • Support for data citation • Science integration guided by EarthScope Science Plan • WInSAR Executive Committee -support for infrastructure

During UNAVCO's strategic planning process, with broad participation of scientists and stakeholders, emphasis has been placed on wide data sharing and on cyberinfrastructure for data access and other data services.

The EarthScope Science Plan and the priorities and planning provided by EarthScope CyberInfrastructure Committee are drivers for understanding the integrative science needs of scientists using data, analysis and visualization tools, and workflows to meet the scientific user requirements.

II. Data Life Cycle for Geodesy

The Data Life Cycle describes in general terms, the typical elements such as data creation, curation, preservation, (re)use, (re)processing, and (re)analysis that occur in scientific data and metadata management and use. UNAVCO's engineering services and Data Center have helped formulate standards for and have actively develop best practices for handling the Data Life Cycle for multiple instrumentation

UNAVCO CyberInfrastructure: Data Life Cycle Support for Geodesy

• Data and metadata standards
• Software tools for format standardization and Q/A
• Data curation, preservation
• Data discovery, search, and access tools
  o For experts
  o For new or non-domain scientists, educators, public
• Software tools to facilitate analysis
• Leverage Data Life Cycle infrastructure for emerging data sets

data sets in the geodesy domain (high precision GPS; borehole strainmeter, seismometer, and pore pressure; laser strainmeter; terrestrial laser scanning) with several thousand supported instruments currently actively feeding data to the UNAVCO Data Center. UNAVCO also manages the purchase or transfer of satellite-based SAR data from space agency repositories to UNAVCO's SAR data and metadata archive on behalf of WInSAR and the GeoEarthScope SAR imagery collection.

NSF's recent focus on data management plans as an integral part of the proposal process recognizes that the initial investment in data collection must be protected and that data are a common resource. Data Centers such as UNAVCO are discipline-focused, and work within a community that is intimately familiar with the data, to develop standards for data and metadata that work. Often the initial needs are for tools that allow for format standardization and initial QA. The software tool TEQC (Translation, Editing, and Quality Checking) was one of the first tools developed by UNAVCO in support of GPS science to do just that. Attention to data and metadata curation and preservation, through practices that maintain and continually recheck the integrity , through planning storage with attention to reliability and robustness, and through leveraging external resources such as cloud-based storage and data processing, UNAVCO supports the curation and preservation aspects of the Data Life Cycle. Format standardization and quality checking are just the beginning of the story. UNAVCO provides supporting software tools for data exploration (time series plotting tools and plate motion calculation are examples), and provides a mechanism to serve community-contributed software tools that solve data manipulation, processing and analysis needs from the UNAVCO website. And along these lines, EarthScope PBO GPS data are processed by a community-vetted University-based analysis center system and this processing has recently been opened to data from non-PBO GPS data sets.

Although some data centers emerge whose audience "knows" where to find services, most centers must continually build and enhance data discovery, search, and access tools so that existing users can make the most of data, or more efficiently access it, and to allow new users to discover and utilize the data. UNAVCO resources are continually directed towards tools to accomplish this, such as the Data Archive Interface for GNSS data and the SAR Spatial Interface for InSAR data. UNAVCO is currently building RESTful web services to replace a previous generation of SOAP/XML web services to enhance data access for GNSS

data and products. Under the direction of the WInSAR Executive Committee, UNAVCO is also building a SAR search and access API based on RESTful web services. Web services are ideally suited to provide Application Programming Interfaces (API) that can form the back end to internal and external tools, to GUIs and to scriptable access tools, opening the search to the broadest set of users.

Having built a highly functional infrastructure to support the Data Life Cycle for GNSS data and for InSAR data puts UNAVCO in a position to advance the design and implementation of analogous infrastructure that supports emerging data sets. In recognition of the need for a widely applicable software infrastructure to provide data cataloging and distribution mechanisms for diverse data sets, UNAVCO is supporting enhancements to RAMADDA, a data and content management system originally developed at Unidata, to quickly implement infrastructure for newer data sets that need these capabilities but don't currently warrant full-on investment in custom infrastructure; an example is the terrestrial laser scanning data sets being generated through UNAVCO's engineering support.

III. Interoperability Priorities

Building interoperability brings choices about where to start to apply resources to add the most value. One focus is expanding the set of users who can find and access data. In the case of high precision GNSS and other geodesy techniques, UNAVCO is one of many existing data centers serving subsets of one or more communities with interest in geodesy techniques. A natural direction for increasing data access is federation of search services. Beginning in the late 1990s, UNAVCO and its peer archives collaborated to develop GPS Seamless Archive Centers. We are currently involved in a NASA ROSES-supported effort with partner archives SOPAC (Scripps Orbit and Permanent Array Center) and CDDIS (Crustal Dynamics Data Information Systems) to develop software to efficiently coordinate federated searches using modern RESTful web services technologies, and to expand the types of data encompassed by the search services. UNAVCO is simultaneously working with international peers and the International GNSS Service to see if these technologies can be adopted beyond US borders.

UNAVCO CyberInfrastructure: Interoperability Priorities

- Tools for expanding data use, re-use :
  - Web services to enable federated searching of analogous archives
  - Web services for search and access for cross-disciplinary data
  - Web services for accessing complementary data sets necessary for processing
- Focused efforts to fill identifiable cyberinfrastructure gaps

In addition to federation to find complete sets of low level data, another high priority cyberinfrastructure task to tackle is cross-disciplinary search and access. Again web services are ideally suited to the task, and UNAVCO has in the past collaborated with IRIS and Lamont Doherty on prototype cross disciplinary search and access OGC web services for GPS, seismic and marine geophysical data sets.

However, because these efforts extend beyond the normal boundaries of domain-specific data centers, it is challenging to gain agreement on which of many possible interactions are the most useful. There is a tremendous amount of work involved in building collaborations and at least at UNAVCO these efforts have not received the serious funding that would be necessary for significant progress to be made.

A third task has to do with providing additional cyberinfrastructure tools to facilitate getting data processed to the derived result when complementary data sets are needed. An example of this is the processing of InSAR data, which requires not just processing software but digital elevation models and atmospheric data for corrections that must be applied. In a separate EarthCube white paper by Meertens and others a cyberinfrastructure based workflow to solve this example is presented.

As the set of cyberinfrastructure tools to support the data use described here is built out, the gaps in available tools for related data sets or for further integration become apparent, and then become the focus for further efforts.

IV. Integration Priorities

Integration of data analyses including complex workflows involving advanced analysis software and visualization tools and utilizing multiple cross-domain data sources ties directly in to the EarthCube goals, while depending heavily on the success of the cyberinfrastructure developments described in the previous sections. Effortless integration requires well-developed analysis and visualization tools, most likely developed by expert domain scientists, as well as the imagination of scientists to clearly put forth the use cases and to identify the tools available. Effortless integration also relies on the best that the cyberinfrastructure experts can put forth in creating workflow engines and accessing multiple large data sets for analyses. The geodetic community, notably through EarthScope, is working to further define the requirements and tools that would facilitate this type of integration and help to realize the goals of the EarthScope Science Plan.

UNAVCO CyberInfrastructure: Integration Priorities
• User requirements gathering • Software tools library for analysis and visualization • Workflow development based on requirements and tools