Dataworks for GNSS: Software for Supporting Data Sharing and Federation of Geodetic Networks

Frances Boler, Charles Meertens, M. Meghan Miller, Stuart Wier, Mike Rost, and James Matykiewicz (All at UNAVCO, Inc., Boulder, Colorado)

IN23B-1734
AGU Fall 2015

Abstract

Continuously-operating Global Navigation Satellite System (GNSS) networks are increasingly being installed globally for a wide variety of science and societal applications. GNSS enables Earth science research in areas including tectonic plate interactions, crustal deformation in response to loading by tectonics, magmatism, water and ice, and the dynamics of water - and thereby energy transfer - in the atmosphere at regional scale. The many individual scientists and organizations that set up GNSS stations globally are often open to sharing data, but lack the resources or expertise to deploy systems and software to manage and curate data and metadata and provide user tools that would support data sharing.

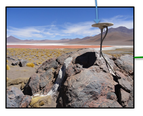
UNAVCO previously gained experience in facilitating data sharing through the NASA-supported development of the Geodesy Seamless Archive Centers open source software. GSAC provides web interfaces and simple web services for data and metadata discovery and access, supports federation of multiple data centers, and simplifies transfer of data and metadata to long-term archives. The NSF supported the dissemination of GSAC to multiple European data centers forming the European Plate Observing System. To expand upon GSAC to provide end-to-end, instrument-to-distribution capability, UNAVCO developed Dataworks for GNSS with NSF funding to the COCONet project, and deployed this software on systems that are now operating as Regional GNSS Data Centers as part of the NSF-funded TLALOCNet and COCONet projects.

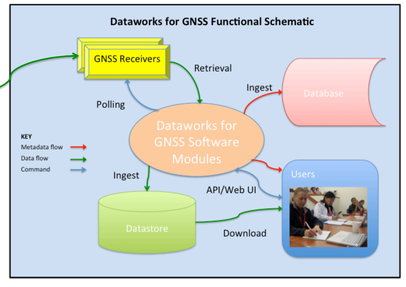
Dataworks consists of software modules written in Python and Java for data acquisition, management and sharing. There are modules for GNSS receiver control and data download, a database schema for metadata, tools for metadata handling, ingest software to manage file metadata, data file management scripts, GSAC, scripts for mirroring station data and metadata from partner GSACs, and extensive software and operator documentation. An Amazon cloud VM image of Dataworks is available that allows standing up a Dataworks-enabled GNSS data center without requiring upfront investment in server hardware. By enabling data creators to organize their data and metadata for sharing, Dataworks helps scientists expand their data curation awareness and responsibility, and enhances data access for all.

Introduction

Dataworks provides data management and distribution software subsystems as open source modules that can be employed by regional GNSS managers for small to medium scale networks (e.g. 10-100 stations).

Recognizing that many organizations operate GNSS stations but do not have the expertise to write their own software systems for the fundamental tasks of GNSS data and metadata management, UNAVCO created Dataworks for GNSS. These software modules are intended to keep the fundamental tasks of handling incoming data, ingesting, metadata storage, and presentation to the users manageable for smaller institutions. In addition to fundamental data management, Dataworks offers specialized functionality such as mirroring and federation among multiple networks.

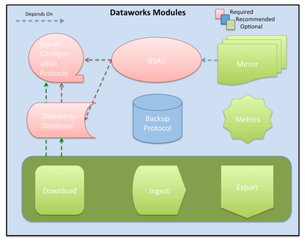




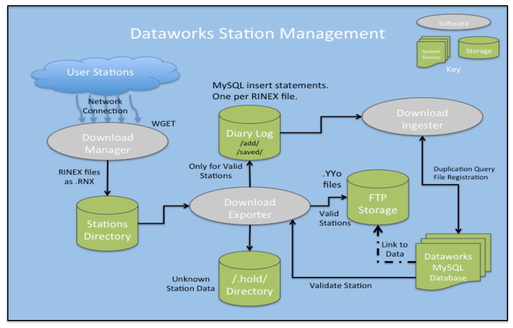
Dataworks Software Modules

Modules include:

- Database schema for MySQL RDBMS for data and metadata management
- Server and Backup Protocols
- GSAC Module for web and API data and metadata access
- Data Acquisition Module (download, ingest, export) for receiver downloading and local data handling
- Data Mirror Module for mirroring of data and metadata from a remote GSAC
- Metrics Module for tracking data downloads and other system metrics

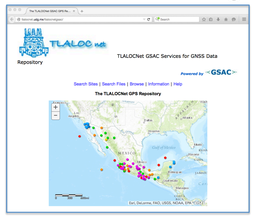


Data Acquisition Module



These modules download files from GNSS receivers (that have internet connections), temporarily storing the files in a local station/data directory, ingesting file information into the Dataworks database, and storing files on a local ftp server for distribution. Ingest and export for data acquisition by other means (e.g. manual download) can also be handled.

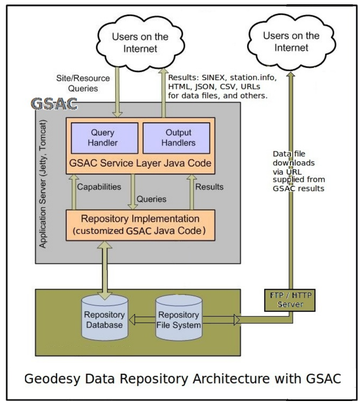
TLALOCNet Regional Data Center



TLALOCNet, a combined atmospheric and tectonic GNSS network in Mexico for the interrogation of climate, atmospheric processes, the earthquake cycle, and tectonic processes of Mexico, leverages NSF and UNAM funding to augment and enhance existing GNSS network stations. When completed, TLALOCNet will span all of Mexico and will link existing GNSS infrastructure in North America and the Caribbean to create a continuous, federated network of networks spanning from Alaska to South America. NSF provided support for the TLALOCNet Data Center, hosted at the Universidad de Guadalajara Mexico, to hold all project data. TLALOCNet uses Dataworks for data and metadata management and distribution.

TLALOCNet Dataworks web site: http://tlalocnet.udg.mx/tlalocnetgsac/

Web User Interface and API

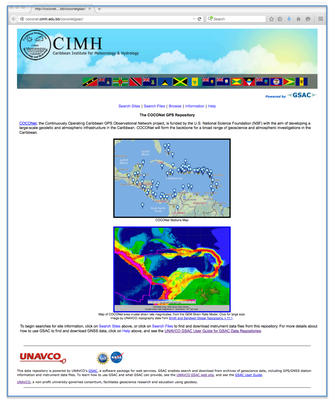


Geodesy Data Repository Architecture with GSAC

Geodesy Seamless Archive Centers (GSAC) software provides a web user interface and an API for metadata and data search and access. Though originally designed for geodesy data types, GSAC is readily adaptable for presenting Earth science data collected at discrete locations and with ongoing data acquisition as files or streams. GSAC utilizes a repository layer and a service layer to identify and present metadata and data services. Numerous output formats for metadata are available and adding formats is simple.

GSAC's web services allow user or machine-to-machine API access, mirroring, and federation.

COCONet Regional Data Centers



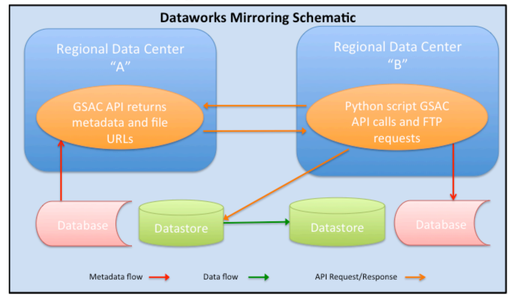
The NSF-funded COCONet project provided support for Dataworks development and for three Regional Data Centers (RDCs), selected via a competitive proposal process, to receive computer servers and Dataworks software, plus modest funding to support operations. The RDCs were awarded to:

- CIMH, Barbados
- INETER, Nicaragua
- SGC, Colombia

RDC technical staff received hands-on training in Boulder on running the servers and Dataworks software. The RDCs provide new or expanded capabilities within the region. A focus on regional data sharing is a benefit of the COCONet RDCs.

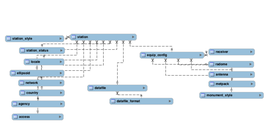
COCONet CIMH Dataworks web site: http://coconet.cimh.edu.bb/coconetgsac/

Mirroring and Federation



GSAC has mirroring and federation capability, which helps to facilitate focused institutional or national data management functions within a context where data sharing can occur among projects that span such boundaries. With federation, each data and metadata collection is kept distinct, yet users can use familiar GSAC queries and web user interface to investigate and access data without needing to know which data center holds any particular part of the overall collection. Administration of the federated GSAC is no more complex than administering an individual GSAC. With mirroring, replication and long-term archiving at another institution such as a long-term archive become simple to manage.

Making Dataworks Easy to Use



Institutions that need tools like Dataworks typically do not have extensive resources or expertise for systems and software installation, support, and troubleshooting. Our goal is to provide useful software that can be easily implemented and supported by small science teams or individuals. Some of the strategies used to accomplish this include:

- Open source software, distributed via GitLab
- Available as an Amazon VM image
- Thorough documentation
- Recorded training sessions available online
- Video tutorials (planned)

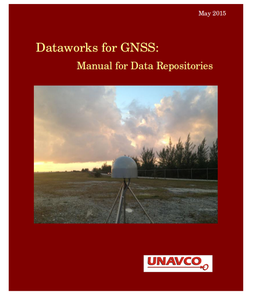
Beyond Dataworks for GNSS

Dataworks for GNSS modules are tailored to the GNSS network data collection scenario. But many field investigations, especially where data are collected at a single spatial observation location, have highly analogous functions to perform. Similarly, these data collection scenarios are led by investigators who are competent to collect and analyze data on their own, but who may not have the needed tools to aid in data sharing, in other words, "the long-tail problem". One way to address this is to provide tools like Dataworks that give the data collectors ownership and control at their institution, yet make data sharing possible. At the same time, these tools can educate about accepted methodologies for defining, curating, distributing, and archiving metadata and data.

To create Dataworks for <fill-in-observation type>

- Define a database schema to capture the metadata
- Create the data acquisition modules
- Adjust the data management file handling scripts
- Simple adjustment of GSAC configuration code may be possible for web discovery, web services, mirroring, and federation

Conclusions



Dataworks for GNSS fills a need within the geodetic community by

- Giving investigators knowledge and tools for data and metadata management
- Facilitating data sharing and preservation
- Allowing broad cross-national networks while preserving local control

For other domains, Dataworks represents a model for addressing the long tail through development of tools tailored to the domain that are accessible to scientists lacking resources and options. Dataworks modules could provide a useful template for other domains.

Acknowledgements: *This work was funded by the National Science Foundation (NSF) and National Aeronautics and Space Administration (NASA) under NSF Cooperative Agreement Nos. EAR-1261833* and *EAR-1042906/9.*

Contacts:
Frances Boler: boler@unavco.org
Charles Meertens: meertens@unavco.org
Meghan Miller: meghan@unavco.org
Stuart Wier: wier@unavco.org
Mike Rost: rost@unavco.org
James Matykiewicz: matykiewicz@unavco.org

Documentation and software access is available on the Dataworks for GNSS web site:
http://www.unavco.org/software/data-management/dataworks/dataworks.html