VLBI Data Ingest Improvements at NASA CDDIS

Taylor Yates[1], Justine Woo[1], Nathan Pollack[1], Jennifer Ash[2], James Roark[2], Sandra Blevins[1], Patrick Michael[3]

Abstract NASA's Crustal Dynamics Data Information System (CDDIS) and the International Very Long Baseline Interferometry (VLBI) Service for Geodesy and Astrometry (IVS) have been collaborating for several years to identify and rectify issues including data and derived product collection completeness and availability. The issues identified include inconsistent quality assurance (QA) across Data Centers, fringe visibilities missing in the archive, latency in resolving data submission issues, and a reliance upon on-premises servers to provide these datasets to the community. In 2021, several improvements to address these issues were made. A new QA architecture has been introduced that utilizes common standards (Data Description Files, DDFs) provided by the IVS. This centralization of QA standards has proven to be vital in improving archive quality and consistency across multiple data centers. SWIN data files contain raw output from the Distributed FX (DiFX) software correlator (Level 1 data) in the Swinburne format [SWIN]. These files are large, compressed directories of the fringe visibilities. Adding SWIN files to the CDDIS archive increases their visibility and use in the community. Additional software has been written that informs data providers when an uploaded file is not recognized, greatly reducing the response time for any anomalies. Cloud deployment of the CDDIS archive will increase data usability via the option to use data in place; therefore, steps are being taken to deploy CDDIS VLBI datasets to be available on Amazon Web Services (AWS) without disrupting active use of the data by the community.

Keywords VLBI, CDDIS, SWIN, DDF, Archive

1 New QA Architecture

The CDDIS ingest processing software now includes new quality analysis (QA) utilizing Data Description Files (DDFs) provided by the IVS which are common to all Data Centers and are source controlled using Git. Unique DDFs for each dataset specify the following:

- Filename scheme
- File destination in the directory
- Product ID for metadata uses
- Data type
- Content type
- Data format
- Validation procedure
- Magic
- Compression type

The CDDIS ingest processing software uses these DDFs as shown in Figure 1.

IVS centralization and control of DDF parameters improves uniformity among the Data Centers by formalizing the requirements for file acceptance and placement.

2 SWIN Data

SWIN Files are compressed directories of observational fringe visibility data in the DiFX format. These are very large (as large as 500 GB per file) and see significant benefit from storage in a centralized archive.

[1] Science Systems and Applications, Inc.
[2] Adnet Systems, Inc.
[3] NASA Goddard Space Flight Center

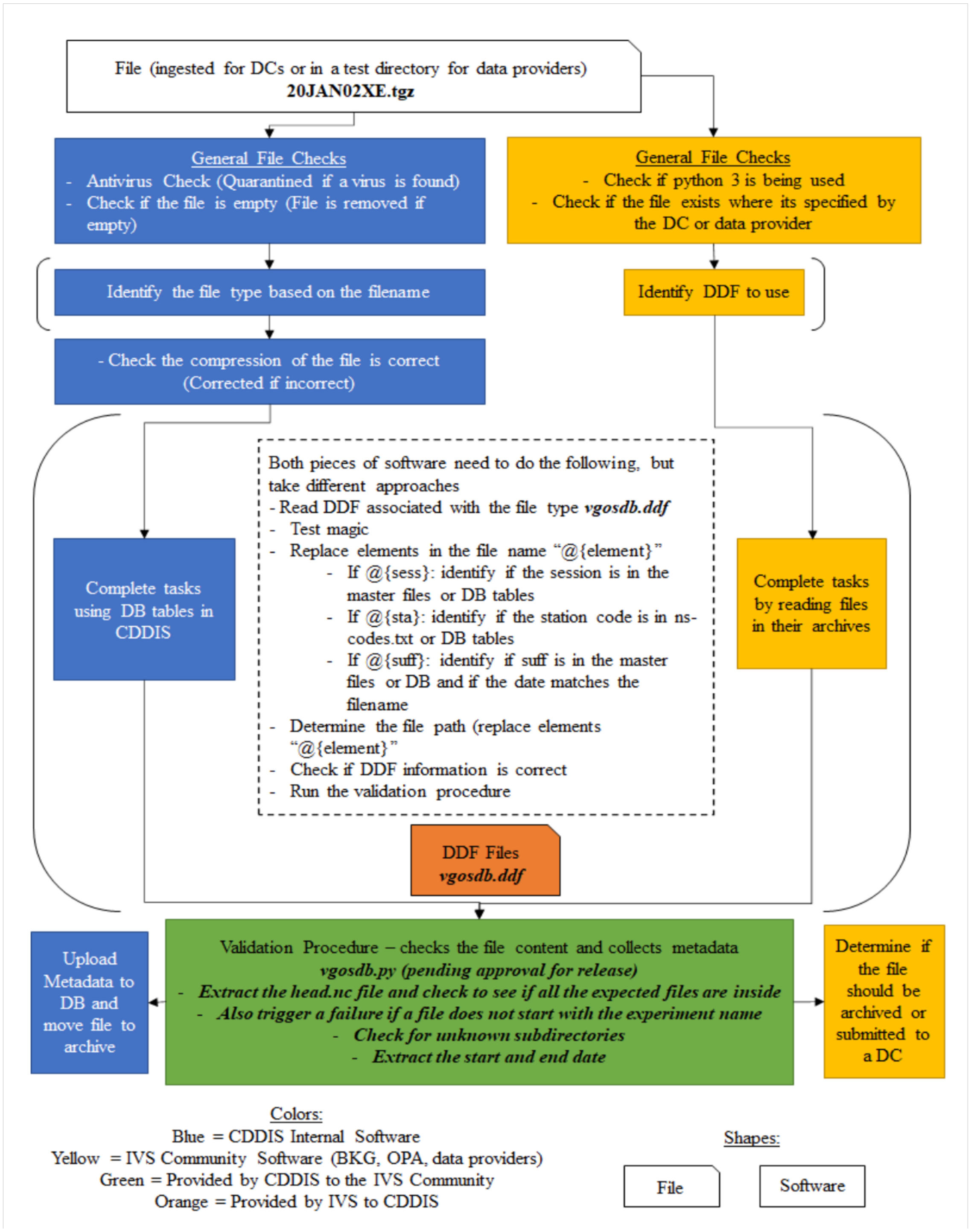
File (ingested for DCs or in a test directory for data providers)
20JAN02XE.tgz
General File Checks
- Antivirus Check (Quarantined if a virus is found)
- Check if the file is empty (File is removed if empty)
General File Checks
- Check if python 3 is being used
- Check if the file exists where its specified by the DC or data provider
Identify the file type based on the filename
Identify DDF to use
- Check the compression of the file is correct (Corrected if incorrect)
Both pieces of software need to do the following, but take different approaches
- Read DDF associated with the file type vgosdb.ddf
- Test magic
- Replace elements in the file name "@{element}"
- If @{sess}: identify if the session is in the master files or DB tables
- If @{sta}: identify if the station code is in ns-codes.txt or DB tables
- If @{suff}: identify if suff is in the master files or DB and if the date matches the filename
- Determine the file path (replace elements "@{element}"
- Check if DDF information is correct
- Run the validation procedure
Complete tasks using DB tables in CDDIS
Complete tasks by reading files in their archives
DDF Files
vgosdb.ddf
Upload Metadata to DB and move file to archive
Validation Procedure – checks the file content and collects metadata
vgosdb.py (pending approval for release)
- Extract the head.nc file and check to see if all the expected files are inside
- Also trigger a failure if a file does not start with the experiment name
- Check for unknown subdirectories
- Extract the start and end date
Determine if the file should be archived or submitted to a DC
Colors:
Blue = CDDIS Internal Software
Yellow = IVS Community Software (BKG, OPA, data providers)
Green = Provided by CDDIS to the IVS Community
Orange = Provided by IVS to CDDIS
Shapes:
File
Software

Fig. 1 CDDIS VLBI QC Overview

SWIN dataset acceptance began in early 2021. During CY21, 2.7 TB of 2021 SWIN data were accepted. Additionally, during this time 3.7 TB of backlog SWIN data from 2017–2020 were accepted into the archive.

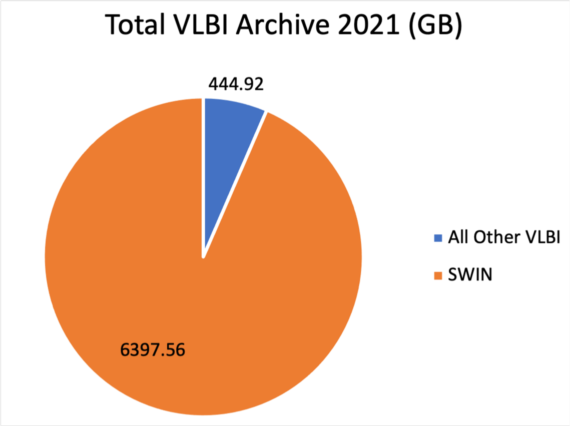

Fig. 2 CDDIS VLBI 2021 archive size.

The introduction of the SWIN dataset increases the CDDIS VLBI archive size by an order of magnitude. In addition, the SWIN dataset itself is expected to grow as more correlators upload SWIN data to the CDDIS archive. This is shown in Figure 3 with data from 2021 and the projection for 2022.

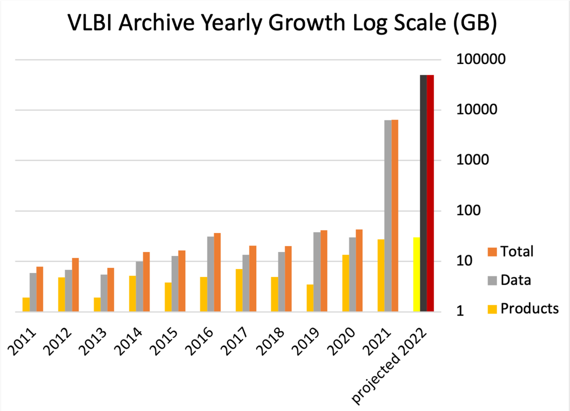

Fig. 3 CDDIS VLBI yearly volume growth.

This growth brings novel challenges to the CDDIS File Ingest System. To resolve these challenges, SWIN uploads use a separate upload web app, ingest script, and partition for archived data. Despite these differences in upload and storage, SWIN availability for users is identical to other VLBI datasets as shown in Figure 4.

The CDDIS SWIN archive is available for public consumption at the URL in Figure 5.

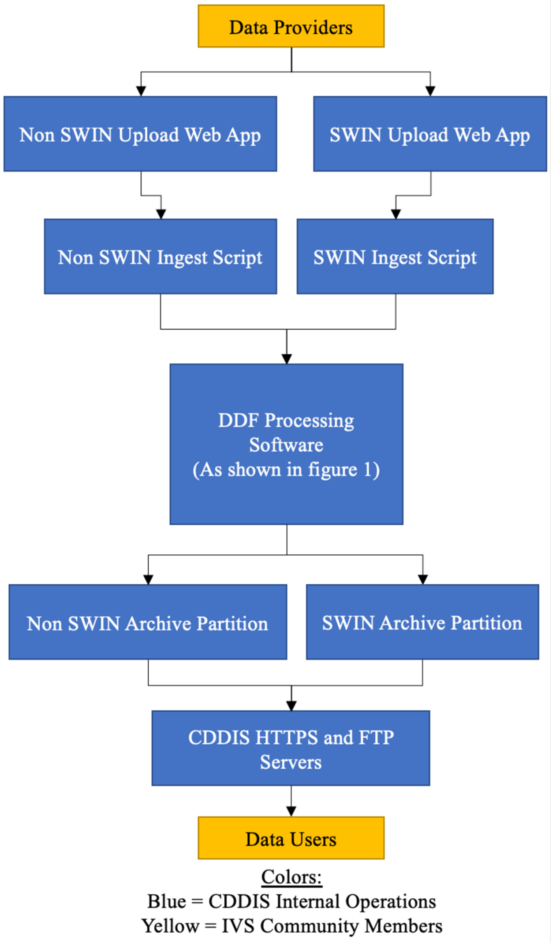
Data Providers
Non SWIN Upload Web App
SWIN Upload Web App
Non SWIN Ingest Script
SWIN Ingest Script
DDF Processing Software (As shown in figure 1)
Non SWIN Archive Partition
SWIN Archive Partition
CDDIS HTTPS and FTP Servers
Data Users
Colors:
Blue = CDDIS Internal Operations
Yellow = IVS Community Members

Fig. 4 CDDIS Ingest/QC/Archive architecture.

Fig. 5 CDDIS SWIN archive HTTPS QR code.

3 Unknown File Error Handling

Files received by CDDIS undergo QA tests to confirm the filetype. Any file that cannot be identified by the QC software is flagged as an ‘Unknown File.’

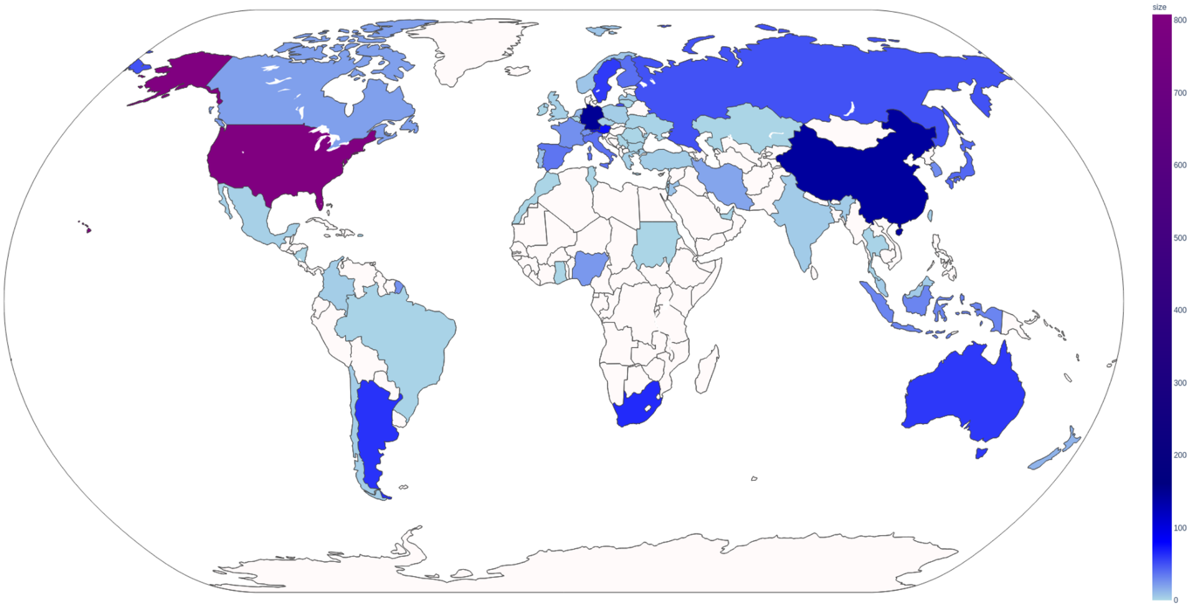

Fig. 6 CDDIS 2021 VLBI unique users.

This can be a point of frustration for uploaders when the file is not available in the archive. To remedy this, software has been introduced to alert providers via email when a file is flagged as unknown.

4 Planned Cloud Deployment

In partnership with other Data Centers in NASA's Earth Science Data and Information System (ESDIS) Project, CDDIS is planning to transition from the current on-premises archive to a cloud-based system, NASA's Earthdata Cloud (EDC).

With data located in AWS S3 buckets, users can choose to either download the data or use it in place. This option bypasses the current requirement to download data from the CDDIS archive to a local machine for use. Unnecessary downloads represent a waste of bandwidth and storage space. Removing the requirement to download data before use will greatly increase the future availability of large VLBI datasets for analysis. With the global userbase that CDDIS services (shown in Figure 6), this would be a significant benefit for the community.

References

1. C. Noll, The Crustal Dynamics Data Information System: A resource to support scientific analysis using space geodesy, Advances in Space Research, Volume 45, Issue 12, 11 June 2010, Pages 1421–1440, ISSN 0273-1177, DOI:10.1016/j.asr.2010.01.018.
2. Deller, A., Tingay, Steven, & Bailes, M.. (2007). DiFX: A Software Correlator for Very Long Baseline Interferometry Using Multiprocessor Computing Environments. Publications of The Astronomical Society of The Pacific, Vol. 119, Issue 853, pp. 318–336, DOI:10.1086/513572.