# VLBI Data Ingest Improvements at NASA CDDIS

Abstract: NASA's Crustal Dynamics Data Information System (CDDIS) and the International VLBI (Very Long Baseline Interferometry) 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 Definition file Formats, or DDFs) provided by the IVS. This centralization of QA standards has proven to be vital in improving archive quality and consistency across the multiple data centers. SWIN data files contain raw output from the Distributed FX (DiFX) software correlator (Level 1 data) in Swinburne format. 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 without disrupting active use of the data by the community.
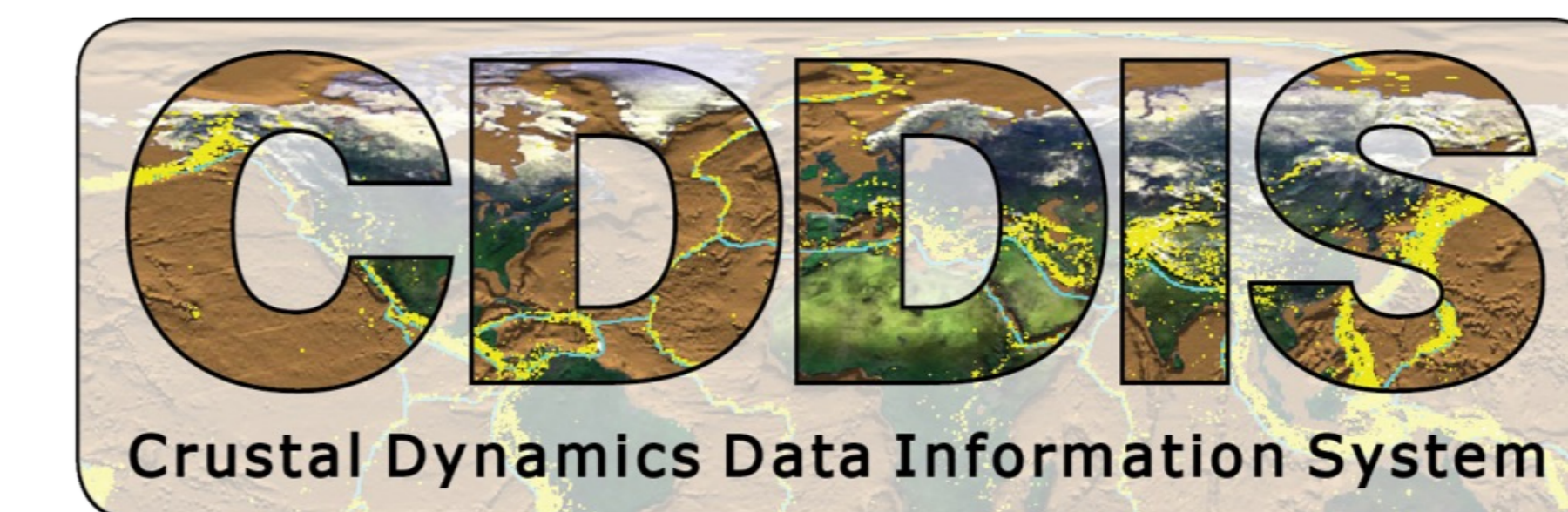
[1]Taylor Yates (taylor.a.yates@nasa.gov)
[1]Justine Woo (justine.y.woo@nasa.gov)
[1]Nathan Pollack (nathan.h.pollack@nasa.gov)
[2]Jennifer Ash (jennifer.ash@nasa.gov)
[2]James Roark (james.h.roark@nasa.gov)
[1]Sandra Blevins (sandra.blevins@nasa.gov)
[3]Patrick Michael (benjamin.p.michael@nasa.gov)

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

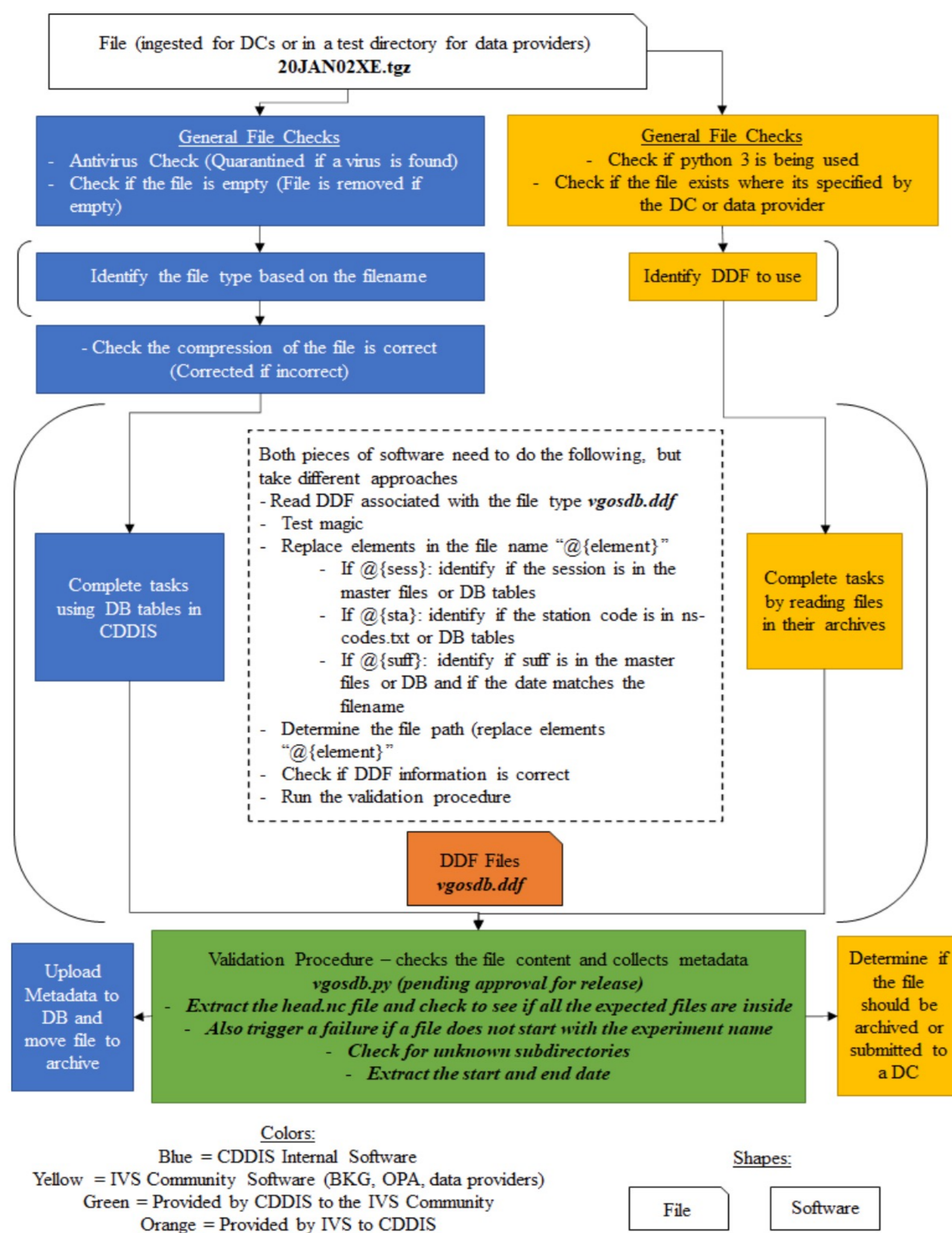IVS centralization and control of DDF parameters improves uniformity among data centers.

Figure 1. CDDIS VLBI QC Overview

SWIN Data

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

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.

The introduction of the SWIN dataset increases the CDDIS VLBI archive 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.
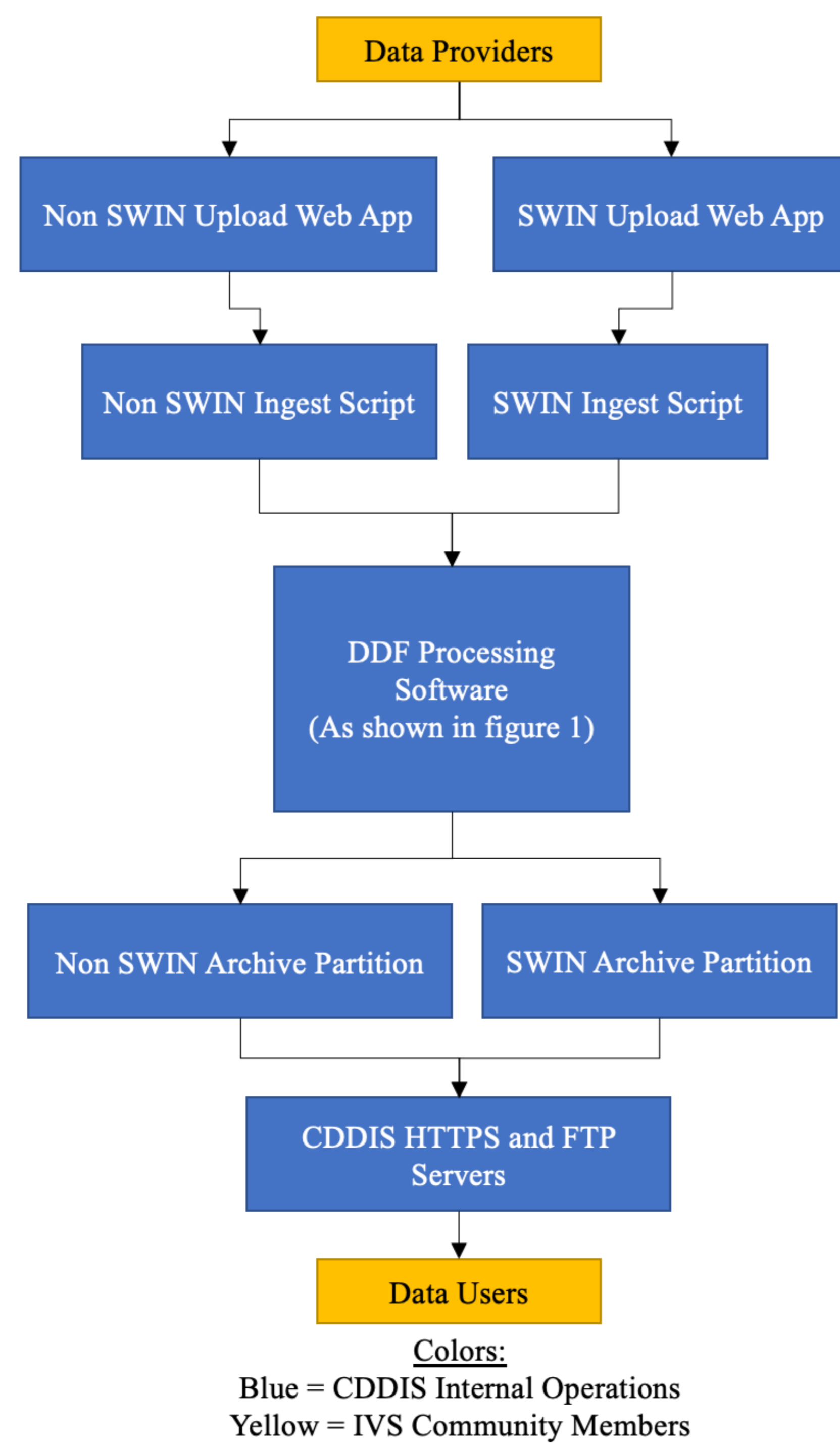

Total VLBI Archive 2021 (GB)

444.92

6397.56

- All Other VLBI
- SWIN
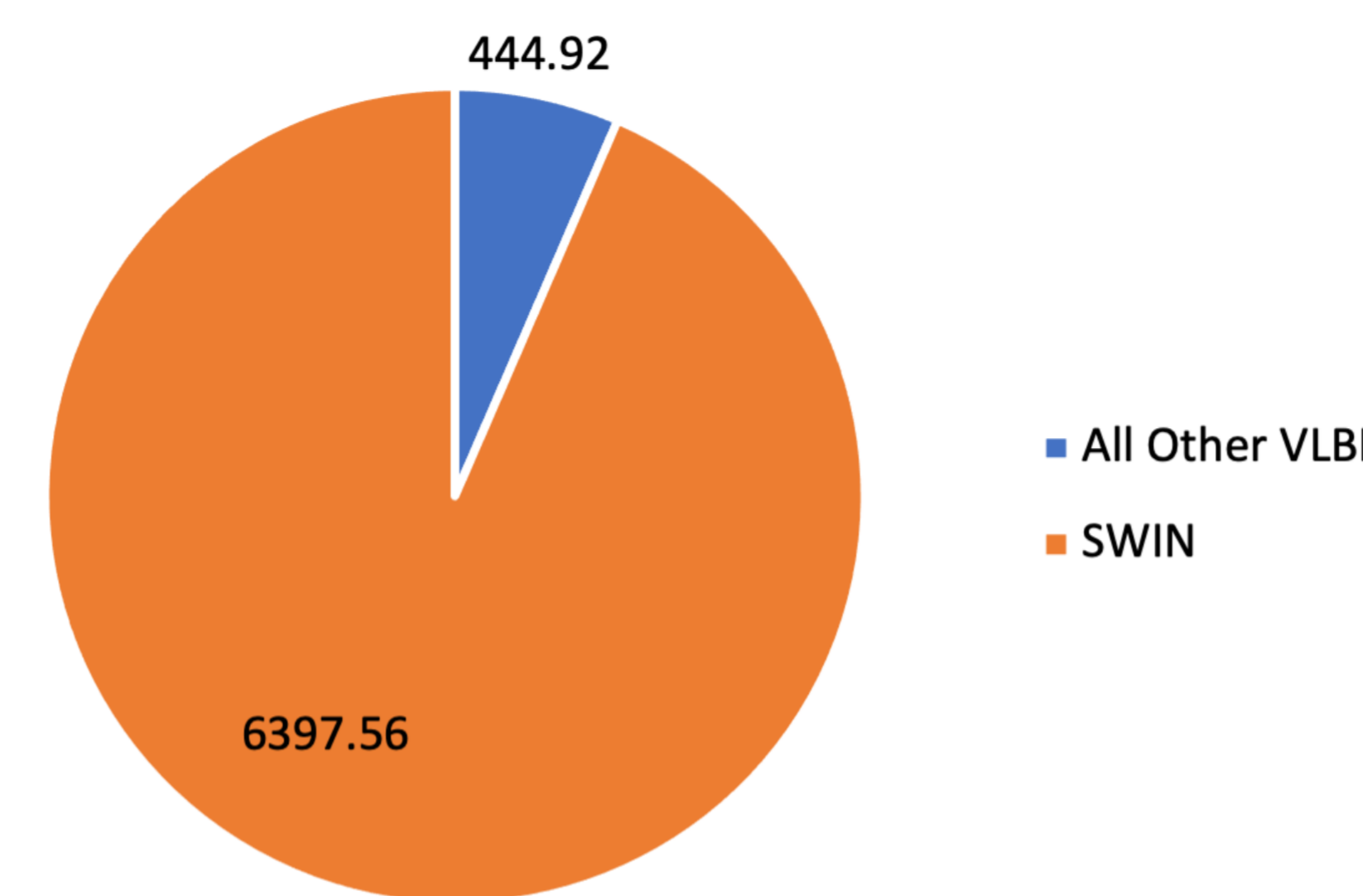
Figure 2. CDDIS VLBI 2021 Archive Size

VLBI Archive Yearly Growth Log Scale (GB)
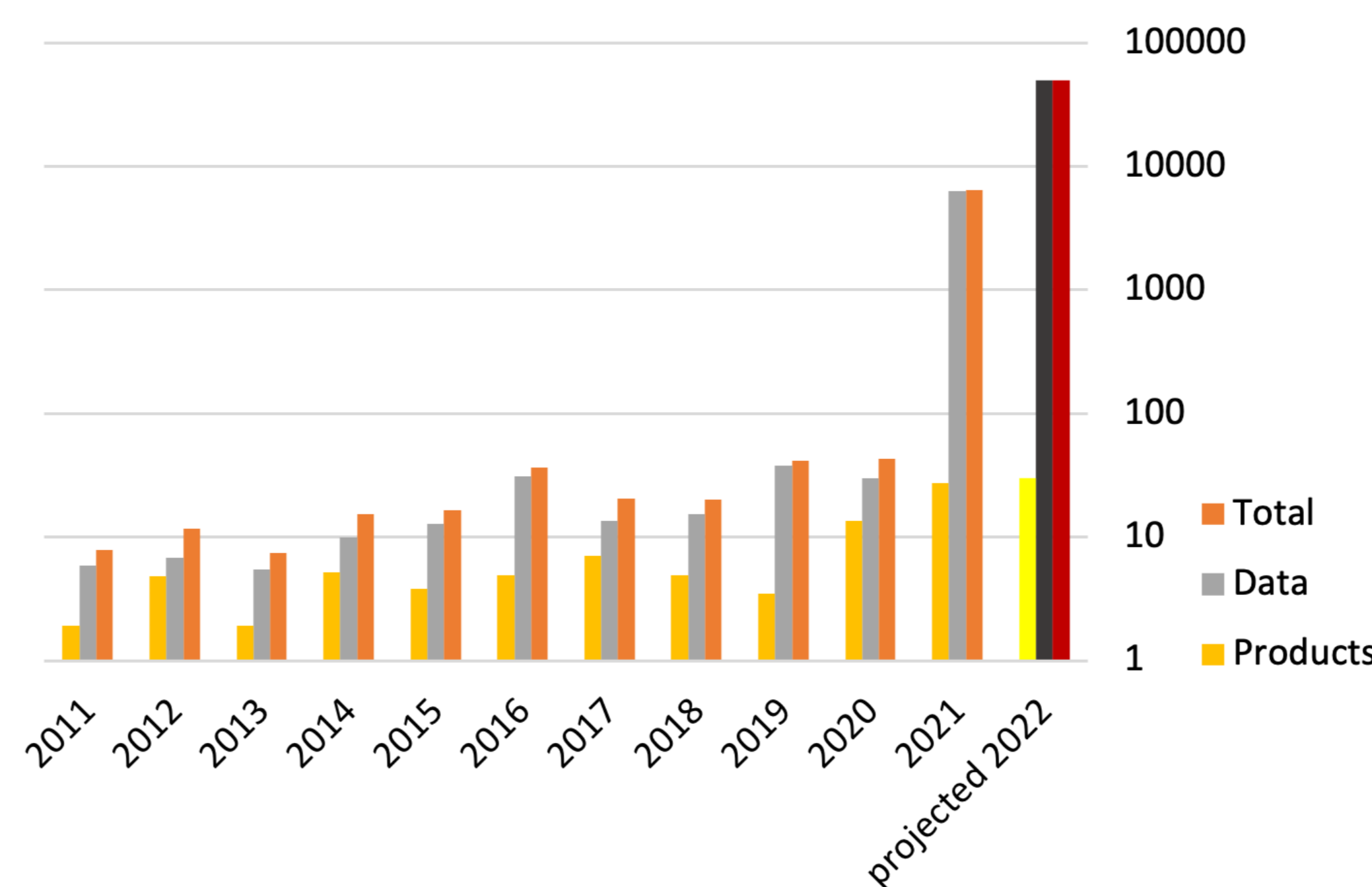
- Total
- Data
- Products

Figure 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.

Colors:
Blue = CDDIS Internal Operations
Yellow = IVS Community Members

Figure 4. CDDIS Ingest/QC/Archive Architecture

Figure 5. CDDIS SWIN Archive HTTPS QR Code
https://cddis.nasa.gov/archive/vlbi/ivsdata/swin/

Unknown File Error Handling

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

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.

Planned Cloud Deployment

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

With data located in Amazon Web Service (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.
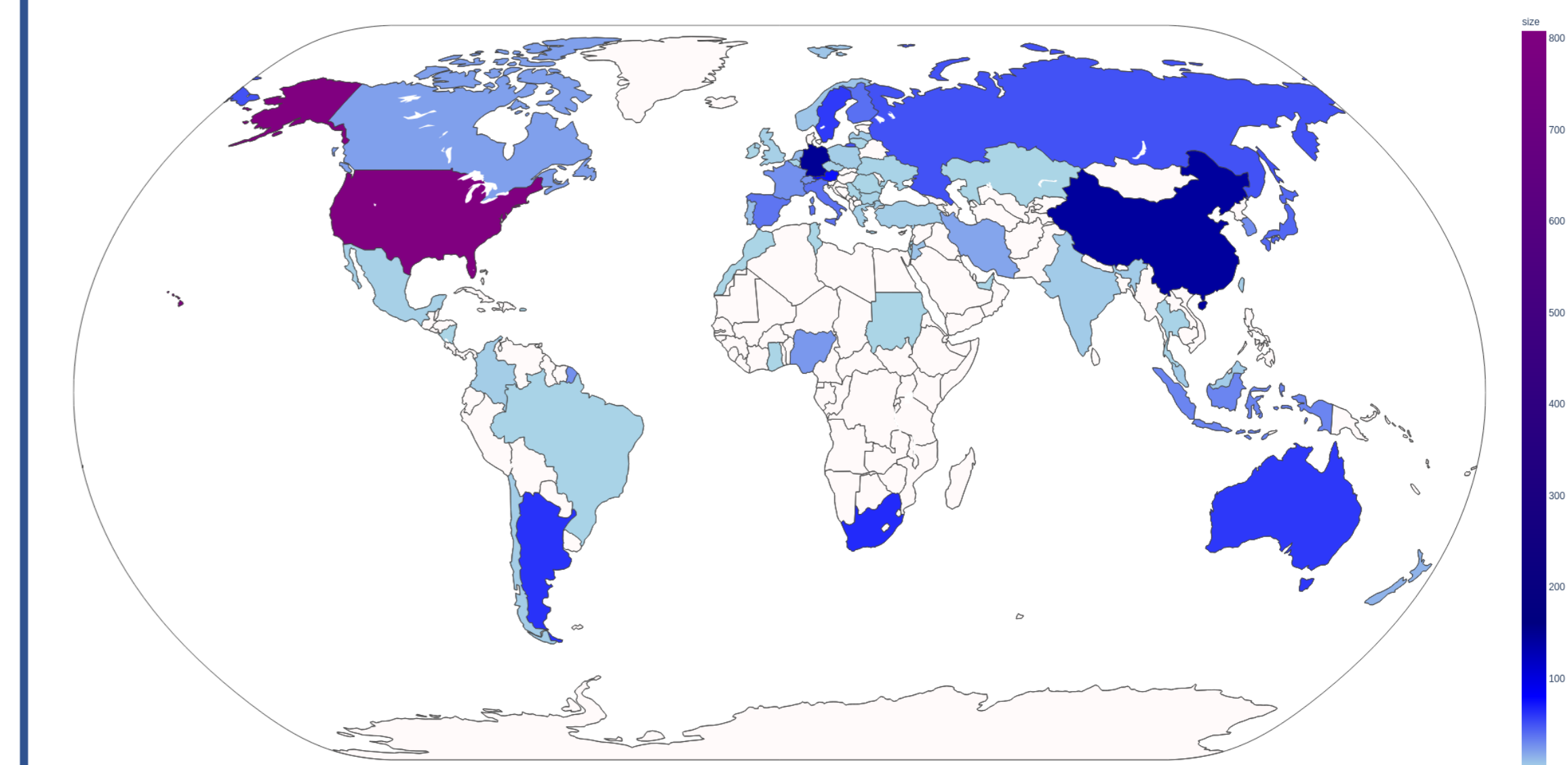

Figure 6. CDDIS 2021 VLBI Unique Users