CGSN Data QA/QC

Sheri N. White, Al Plueddemann, Andrew Reed

Peter Brickley, Stephanie Petillo

Jennifer Batryn, Collin Dobson, James Kuo, John Lund, Allen Smith, Rebecca Travis

CGSN Data Team

Data Team Leads

- **Sheri N. White** – Systems Engineering

- **Al Plueddemann** – PI/PS

- **Andrew Reed** – Data QA/QC*

Data Team Support

- **Peter Brickley** – Operations
- **Stephanie Petillo** – Software

* only Full-Time Data Team member

Data Team Support (cont'd)

- **Jennifer Batryn** – Instrumentation

- **Allen Smith** – Instrumentation

- **Collin Dobson** – Operations, gliders

- **James Kuo** – Subsurface Moorings

- **John Lund** – Subsurface Moorings

- **Rebecca Travis** – Config. Management

Spending ~1.5 FTE effort on Data QA/QC in OOI 2.0 (Oct 2018-Apr 2019)

CGSN Data Team Responsibilities

- **Asset Management**
  - Entering required metadata (cal, deployment info) into OOINet
- **Data Ingestion/Availability**
  - Ingesting/re-ingesting data into OOINet
  - Ensuring data flow from platform to shore to science users
- **Data QC**
  - Identifying/resolving issues with instruments/data
  - Implementing automated and HITL data QC
- **Annotation**
  - Annotating data in OOINet for instrument/data issues, metadata changes

Asset Management

- Improving asset management processes
  - Double-checking all entered metadata through GitHub
    - Using a dedicated fork and branches for calibrations and deployments
  - Developing automated processes for capturing calibration coefficients
    - To prevent manual entry errors, and speed up the process
  - Documenting SOPs

- Added or Uploaded 442 CSV files in OOI 2.0
  - calibrations, new deployments, updates from metadata review

- Critical Metadata review of all past metadata in OOINet

Data Ingestion/Availability

Data ingestions

- OOI 1.0 known ingestion backlog complete: 25 streams ingested
  - Will perform further ingestions as dictated by Redmine ticket review
- OOI 2.0 data ingestions: 741 streams ingested
- OOI 2.0 data ingestions ongoing: 95% of available ingestions completed (47 recovered ingestions pending)

Data availability

- Adapting OSU ERDDAP scripts to assess data availability

Data QC

- Operations and Data Monitoring
  - Implemented formal weekly CGSN field operations review, assignment of engineering & data actions, and updating OMS++ infrastructure and instrument alarming

- Addressing issues via OOI Data Team working groups
  - QARTOD, ADCP, Communications, Ticketing, Water Sampling, etc.

- Working high-nail instrument issues (common failures, Redmine tickets)
  - Helpdesk tickets assigned to CGSN in OOI 2.0: 4 tickets closed, 9 tickets open
  - Operations & Hardware tickets in OOI 2.0: 36 closed, 5 open

Annotations

- Tied into and fall out from other Data Team efforts
  - Monitoring operations of platforms
  - Monitoring data quality of instruments
  - Changes to Critical Metadata that affect data

- Working on processes for notifying the community about annotations
  - Data Communications Working Group working on this

- Number of annotations created or updated in 2.0
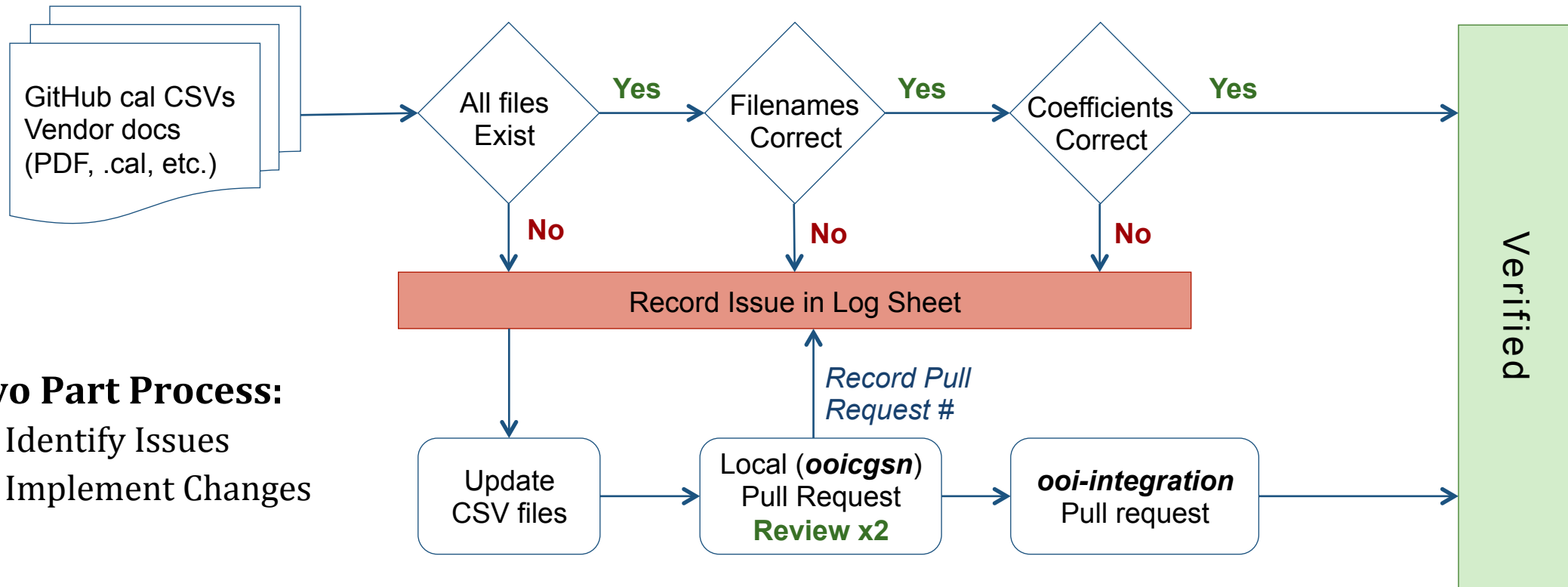 - 61 created/updated   (88+ more identified to be done)

CGSN Critical Metadata Review

Scope of work

- Types of metadata to be reviewed and compared to other documentation (vendor sheets, cruise logs, etc.)
  - Calibration Coefficients – 2492+ calibration files, 28,394+ calibration coefficients
  - Platform Deployments
    - Moorings – 27 moorings, 147 total deployments
    - Gliders – 52 Gliders, 135 total deployments
    - AUVs – 2 AUVs, 15 total deployments
  - Assets
    - 1062 Instruments in inventory (sensor_bulk_load)
    - ~300 Mooring control components (eng_bulk_load)

CGSN Critical Metadata Review Process

Compare GitHub calibration CSVs with vendor documentation

GitHub cal CSVs
Vendor docs
(PDF, .cal, etc.)

All files Exist — **Yes** → Filenames Correct — **Yes** → Coefficients Correct — **Yes** → Verified

No ↓   **No** ↓   **No** ↓

Record Issue in Log Sheet

Two Part Process:
1. Identify Issues
2. Implement Changes

Update CSV files → Local (*ooicgsn*) Pull Request **Review x2** → *ooi-integration* Pull request → Verified

Record Pull Request #

CGSN Critical Metadata Review Process

Verify platform deployment details and component serial numbers

GitHub
deployment CSVs
Config Sheets &
Mooring Logs

Deploy
Date / Time
Correct
— **Yes** →
Lat / Lon
Depth
Correct
— **Yes** →
All S/Ns
Correct
— **Yes** →
Verified

No ↓ **No** ↓ **No** ↓

Record Issue in Log Sheet

GitHub CSVs:
  bulk_load
  Vessel
  CruiseInformation
  Deploy

Update
CSV files
→
Local (*ooicgsn*)
Pull Request
Review x2
→
ooi-integration
Pull request

Record Pull Request #
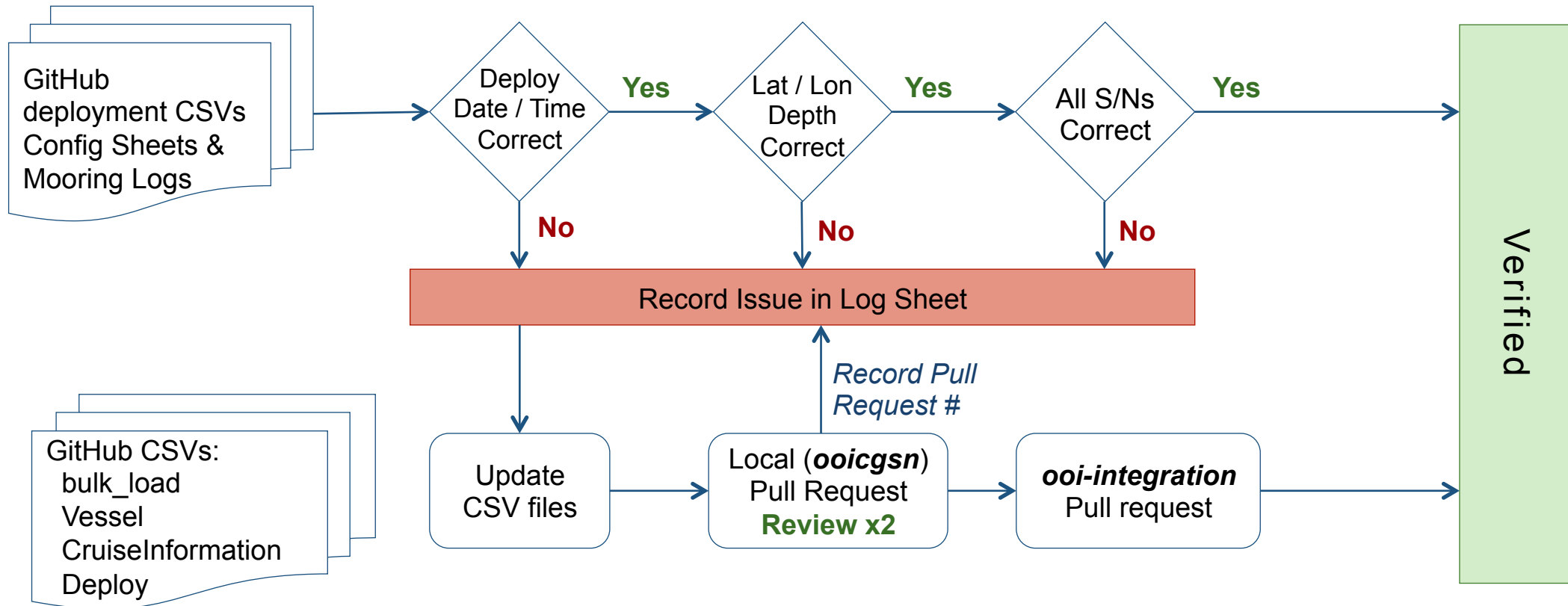
CGSN Critical Metadata Review Status

- Efforts include creation of automated tools for metadata checks
  - Necessary due to volume of instrumentation (e.g., hundreds of CTMOs)
  - Scripts will be used both for current Metadata Review and on-going Asset Management updates (to streamline and reduce manual errors)

- Need for metadata review: over half of files show inconsistencies
  - Mostly issues related to consistency of naming files (not data-related)

- Impact to data users: <10% of calibration coeff show errors

- Aiming to have Critical Metadata Review complete by end of PY1
  - Currently ~50% of cal files reviewed

CGSN Accomplishments To Date

- Full-time CGSN Data QC hire in November 2018 (Andrew Reed)

- Attended OOI Data Team Workshop at UW in January 2019

- Completed re-ingestion of backlog from OOI 1.0

- Participated in creation of a common format/spreadsheet for OOI Water Sampling data across the program
  - Continuing implementation of that for all past data
  - Suggested ingestion of water sampling data into OOINet

- Successful Asset Management updates and Data Ingestion for:
  - Pioneer 11, 11a, 11b, 11c; Southern Ocean 5; Pioneer 12, 12a cruises

CGSN Plan Going Forward

- Weekly meetings to identify work and tasking

- Continuing critical metadata review

- Evolution of OMS++ system to improve mooring and data monitoring (alerts & alarms, data visualization)

- Engaging SMEs for to confirm/determine best practices and data review methodology

- Continue to develop and implement automated processes for metadata and data review

- Continue OOI Data Team work on Program-wide issues/processes