OCEAN OBSERVATORIES INITIATIVE

OOI Data Delivery Systems: Present and Future

Wednesday, October 26th, 2022

Jeffrey Glatstein
Senior Manager of Cyberinfrastructure

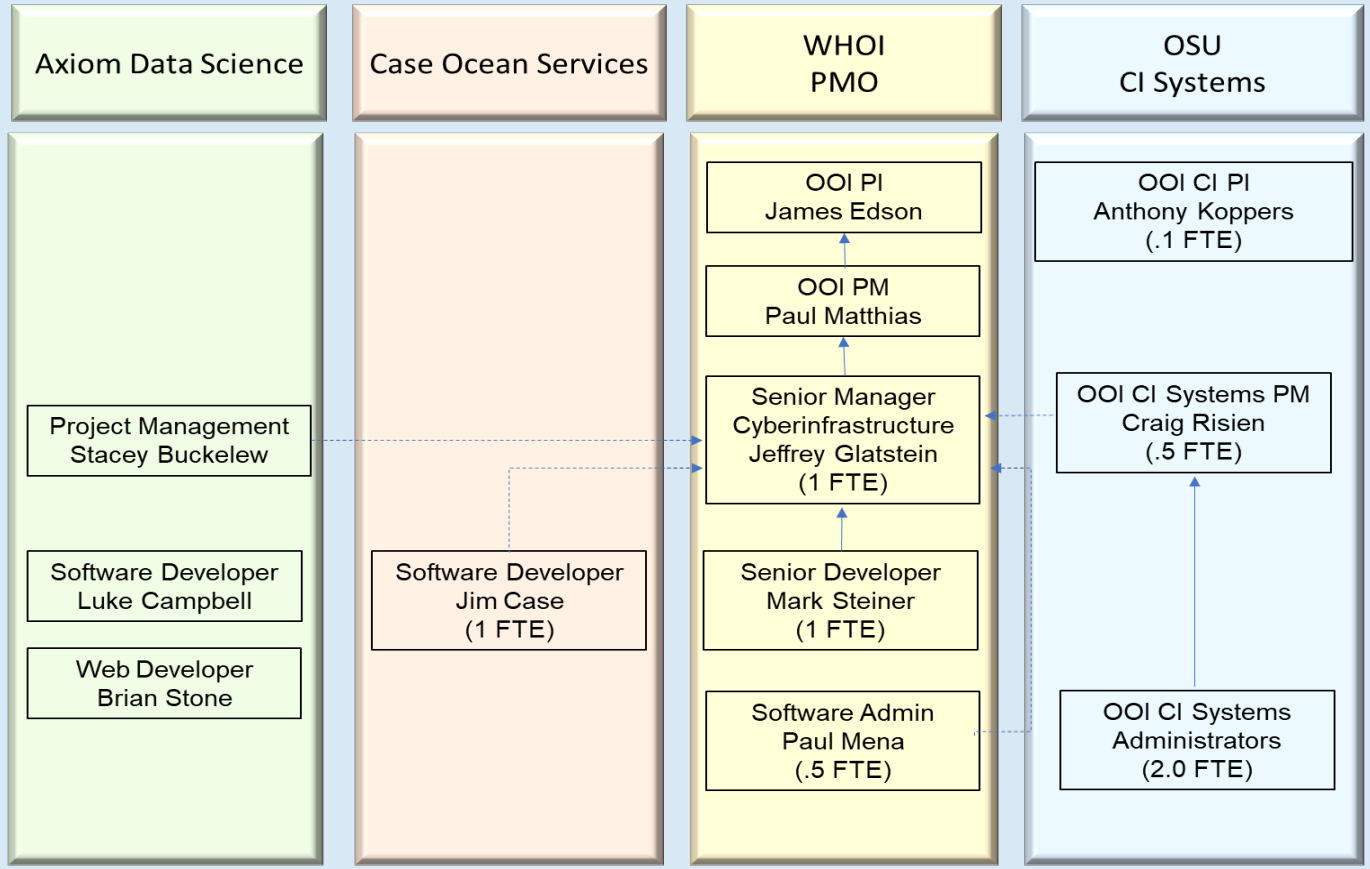
OCEANOBSERVATORIES.ORG

Agenda

- Resources
- Data Processing Components
- Significant Projects and Impacts
- Objectives for PYV
- Cloud Strategy
- Five Year Roadmap Concept

CI Resources

- Senior Manager of Cyberinfrastructure and Data Delivery Manager (PMO) - responsible for all aspects of the OOI Cyber Infrastructure (strategy, budget, and execution), data delivery (including UX), and execution of a QA/QC program.
- CI Systems Project Manager (OSU) – responsible for day-to-day operations, including prioritization of operational tasks, management of Systems Administrators, budgetary execution for purchases and renewals, executing on strategic priorities, and development and submission of required reports.
- Systems Administrators (OSU) – responsible for the monitoring and maintenance of the OOI CI hardware and network infrastructure.
- Software Administrator (PMO – WHOI Information Services) – responsible for building and migrating the OOI software, monitoring, GitHub ownership, and 3rd party applications.
- Lead Software Engineer (PMO) – responsible for uFrame and data ingestion components and tasked with reviewing other developer's designs and code.
- Software Developer (Case Ocean Services) – responsible for maintaining and retiring the legacy Data Portal, web services supporting Data Explorer, multi-media processing and asset metadata delivery.
- Project Manager (Axiom Data Sciences) – responsible for coordination and management of Axiom resources developing the Data Explorer tool.
- Software Developer (Axiom Data Sciences) – responsible for data ingestion and interface processes into the Data Explorer tool,
- Web Developer (Axiom Data Sciences) – responsible for the UI for the Data Explorer tool.



Data Processing Components – High Level

- Data Processing
  - Databases (Cassandra and PostgreSQL)
  - Edex
  - Data Ingestion
    - Ingest engine
    - Data parsers
    - Queues by delivery method (cabled, telemetered, recovered and playback)
  - Data Delivery
    - StreamEngine/M2M
      - Preload database
      - ION functions
    - QARTOD
- Data Discovery
  - Data Explorer
  - Data Portal (OOINET)
  - Thredds ‘Gold Server’
  - Raw Data Server
  - Jupyter Hub (Alpha)

Significant Projects and Impacts to Date

- Performance
  - Cassandra database tuning and cluster size increase
  - New architecture – virtualization of uFrame (part of data center move)
  - OOI software and components upgrade – edex, 3rd party software and databases
  - Implementation of StreamEngine query governor
  - Removal of worst-case scenario data retrieval as default on OOINET
  - Maintain server of precalculated datasets
- Data accuracy and FAIR
  - Implementation of QARTOD data quality code – Gross range and Climatology
  - Data Maintenance – ability to purge and replay data by time range
  - Asset management data review
  - Preload database corrections for CF compliance
  - StreamEngine aggregation tuning
  - Resolution of data quality tickets

Significant Projects and Impacts to Date

- User Experience
  - Adjustment of OOINET interface utilizing user feedback
  - Implementation of Data Explorer with user driven use cases
  - Move to precalculated data sets (calculate on demand still available)
  - Established user feedback loops (e.g. Discourse)
- Efficiency and Effectiveness
  - Data back-ups – tape, cloud and built-in redundancy
  - Improved Cybersecurity with Trusted CI relationship and system vulnerability scanning
  - Monitoring – effectively tells CI about issues prior to the user
  - Advanced communications plans and fostered environment of cooperation

Objectives for PYV

- Stream Engine re-architecture
  - Upgrade to Python 3 (SE Code and all ION functions)
  - 30+ requirements – Reporting across reference designators, .zarr file support, multi-level co-located instrument data
  - Data request management – load balancing, request management routes to cancel requests
- Data Explorer
  - Completion of full resolution data visualization
  - Expansion of media server to include HD video, Hydrophone and streaming data
  - Data Explorer operational training to OOI development and operational resources
  - Further reingestion automation and reporting
  - ZPLS and AUV data availability
  - Addition of remaining scientific data
- Compute in place - Jupyter Hub beta release
- Asset management – Roundabout development

Objectives for PYV

- Data Accuracy and FAIR
  - Continue to target data quality tickets
  - Continue QARTOD support and development of test and tools
  - Continue to support preload database analysis and adjustments
  - Continue FAIR data standards tuning (Jupyter HUB, Preload database adjustments)
- Performance
  - Query performance analysis
  - Integration of new processing and storage resources
- Operational
  - Cloud storage transfer to TACC
  - NCEI data archival
  - Dev-ops, Monitoring and improved efficiency of releases
  - Database replication
  - Disaster recovery scenario exercises

Objectives for PYV

- Strategic
  - ERDDAP tuning and replacement evaluation
  - Deliver Digital Object Identifiers (DOI) recommendations for policy and approach
  - Analysis of Alternatives for Alfresco, Confluence and Jira
  - Evaluate options to reduce or eliminate the Cassandra/PostgreSQL database footprint
  - Continued cloud analysis

Cloud Strategy

- Continue cloud analysis on a yearly basis – produce white paper
- Research has shown moving OOI computing infrastructure to the cloud is not cost effective and provides little benefit
- The benefit of cloud to OOI is in the democratization of compute facilities and access – efforts should be directed to this end goal
  - Continue evaluation of cloud friendly files types (.zarr already planned)
  - Compute in place – Jupyter Hub
  - Evaluate cloud implementation of OPeNDAP

Five Year Roadmap Concept

- Rolling 5 years
- Currently 80+ entries
- Line item can have various designations such as “in process” or “abandoned”
- High level view of line items
  - Replacement of Cent OS and upgrade of OOI software stack
  - Placement of engineering data and new Data Explorer persona analysis
  - Document management
  - Data Lake (Data Lake House) model connection to cloud computing
  - Raw data discovery interface
  - System storage cap and discoverability
  - System and data status page
  - Migrate off of Python

OCEAN
OBSERVATORIES
INITIATIVE

Questions?