NTL Data Acquisition by GCOOS: Implementation Plan

Version Control

2019-10-07	0.1	Initial draft for discussion (F. Gayanilo, L. Belabbassi, B. Kirkpatrick)

The data from the industry sector in the Gulf of Mexico, which include more than just ADCP data products, are currently being captured and distributed by NDBC. The responsibility to capture and disseminate the NTL data from 128 platforms (54 active platforms) will soon be transferred to GCOOS. The number of active and inactive industry platforms fluctuates. The following are the tasks required to initiate the transfer of responsibility and to sustain the operations.

The information system to be developed and deployed will be referred to here as ***GCOOS-NTL-DAS*** (Data Acquisition System). Figure 1 is the *Data Flow Diagram* (DFD) of the proposed information system, ***GCOOS-NTL-DAS.*** The initial phases of the project will involve a simulated transfer of data from NDBC, and in the later phase of the project (Task 4.4), the data communication from NDBC will be ignored and data are expected to come directly from data sources (projected in 6 months after the project has started).

Task 1. Project Scoping

The NTL industry data package delivery and the NDBC product services will need to undergo detailed re-scoping with NDBC. A 1-day workshop will be scheduled to ensure that all the parameters of the operations from data ingestion to data delivery are understood.

> Note that NDBC has already provided the data links and other documents to GCOOS to pre-analyze the tasks. These materials will be used as the basis of the 1-day workshop with NDBC.

Task 2. System Design and Development

The project scoping in Task 1 will be used as the input to design an information system to be responsive to the demand of the tasks at hand and to be able to sustain the operations.

Task 2.1. Functional requirements

The functional requirements from data acquisition to the point of distribution will be listed and documented as part of the design document. These will include requirements to receive the data, storing the submitted data, parsing, ingestion, curation, reporting, and distribution.

Initial meetings with NDBC reveals that the following are known data transport scenario, and will be considered in the system design document:

- Most common, data is submitted to NTL designated repository as soon as the data is available (most ideal);
- Consolidated data is submitted to NTL designated repository, not in realtime (daily, monthly, etc.; delayed mode);

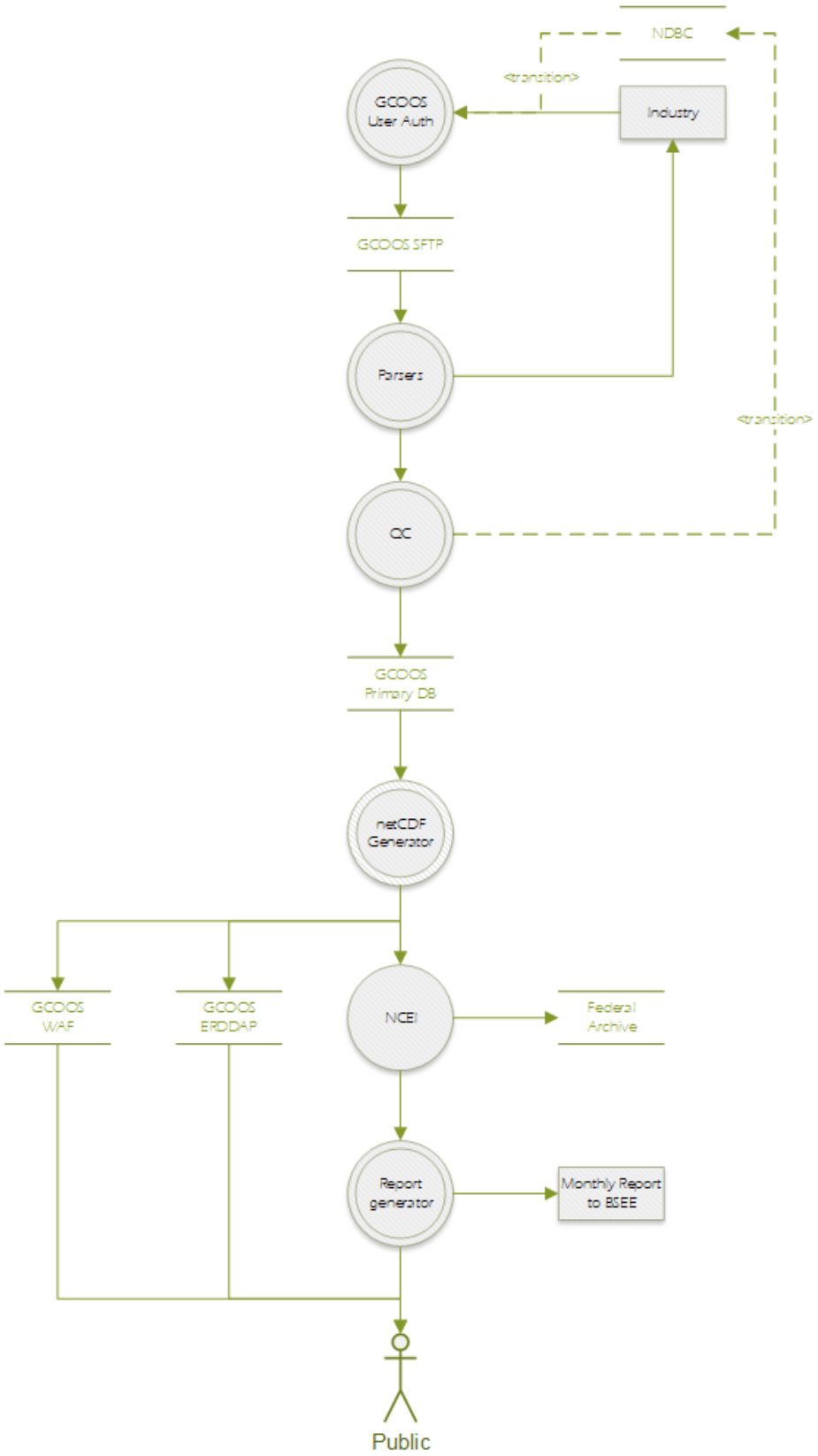

*Figure 1. Data Flow Diagram for the proposed **GCOOS-NTL-DAS** to acquire NTL data.*

- Data are not always submitted in the proper temporal stamp requiring a resubmission, and in cases where communication breaks down at the data source, data gaps are not submitted to the NTL designated repository;
- Data are often submitted more than once, requiring an overwrite of previously saved data (slows down processing);
- Data are transported largely in binary forms requiring a module to parse with a structured header/leader attached;
- Data structures are modified by the data provider without notice requiring the acquisition of 3rd party modules and maintaining several versions (OS dependencies is not known); and
- Data are not quality checked at the source.

The industry data does not only provide data from ADCP deployments (measuring currents at depth) but also willing to share all other IOOS observable parameters (e.g., air temperature, barometric air pressure, winds, water temperature, turbidity, waves).

Task 2.2. Timeline and Deliverables

Immediately following the completion of Task 2.1., timeline and deliverables will be re-visited and adjusted accordingly. The desired timeline will include the following:

- SFTP site should be ready to receive data in month-2 of operations;
- Data WAF is available for raw data in month-3;
- Data in NetCDF files should be available in month-6 in the WAF and GCOOS ERDDAP;
- User interfaces for interactive browsing and querying of the data in month-8; and
- D/R facility is fully operational in month-10.

Task 3. SFTP Installation and Deployment

Secure File Transfer Protocol (SFTP) will be the mode of data transport from the data sources. Data providers have established their automated system pointing to the NTL-designated repository, and the only change will be the URL of the SFTP.

Task 3.1. Install SFTP Server

GCOOS will install an SFTP server only to manage and transport NTL data. An alternate, offsite, SFTP server will also be installed (see Task 6). The new NTL-designated data repository site will be sftp://ntl.gcoos.org.

Task 3.2. Setup authorization and authentication

A guideline on account acquisition for SFTP will be drafted by GCOOS in close collaboration with NDBC following existing arrangements made by the industry with NDBC. However, few more optional parameters will be requested as the need arises to ensure proper data provenance (e.g., originating IP).

The SFTP server will be configured to recognize servers registered on the system using originating IP addresses. A series of online meetings will be organized to ensure that the parties involved are aware of GCOOS policies (see also, Task 5.1).

Task 4. System Development and Deployment

This is the main task for year 1 of operation. New modules will need to be developed and interfaced with existing modules.

Task 4.1. Review of Functional Specifications

The system functional requirements for ***GCOOS-NTL-DAS*** will again be reviewed for any last minutes requests and changes to the design. It should be noted here that the majority of the modules required to operationalize ***GCOOS-NTL-DAS*** are already in place. Focus will be made on the following: (1) data acquisition, i.e., as data is transferred to GCOOS SFTP site, (2) parsing the data received, (2) monitoring of NTL data submission, (3) alert systems, (4) D/R for ***GCOOS-NTL-DAS***, and (5) reporting requirements to BSEE and the public.

Modular/tiered design parameters will be identified in this phase of the project.

Task 4.2. System Coding

The additional modules needed for ***GCOOS-NTL-DAS*** will be developed using Python and Python-related technology stacks. All codes will be deposited to the GCOOS GitHub repository and it will adhere to existing open-source policies of GCOOS. Agile development approach will be employed for proper phased development of the additional modules.

Task 4.3. Unit and System Test

GCOOS-NTL-DAS modules will be unit tested, and a week will be scheduled for system testing. NDBC and BSEE will be invited to test the new modules. During the testing phase, all procedures and actions taken will be recorded and will be used to configure regression testing to new or updated modules prior to the publication (i.e., transfer to production server) of the modules.

Task 4.4. Parallel Run with NDBC

In this phase of the work, NDBC will begin to push data to GCOOS SFTP, and GCOOS will start processing the data (Figure 1). Both parties will evaluate the procedure and make the necessary corrections as the need arises. This phase will include the test publication of the data on ERDDAP and the GCOOS WAF.

This Task is expected to last for about 3 months or until all issues resolved. All issues will be noted on the GCOOS GitHub for proper tracking and management.

Task 5. Maintenance and Administration

Task 5.1. System documentation

The project will develop an online site (MkDocs, https://www.mkdocs.org/, is preferred) to document on how to use the online facilities effectively, best practices in file naming, data transport, checking for data completeness, and local-site data management.

Task 5.2. Code Repository and Management

As stated in Task 4.2, all codes will be published via the GCOOS GitHub account. The repository will be backed-up once a month when necessary. Pull and merge requests will be managed by GCOOS staff, but users will be encouraged to request a pull when applicable.

As new data types are introduced or a change in the data structures are made, modules to address the change will be developed, tested and deployed.

Task 5.3. Issue Tracking

To maintain transparency, publicly-accessible GitHub pages will be published to allow inquiries to be tracked and monitored. GitHub resources at Texas A&M University and Atlassian Jira will be explored by GCOOS.

GCOOS reserves the right to delete posted comments and inquiries that are not NTL-data-related. In the outset of the project, GCOOS will draft communication guidelines to ensure proper handling of user (data providers, data users) requests. All oral communications related to ***GCOOS-NTL-DAS*** will be recorded and complied with all other written communications.

Task 5.4. Reporting Functions

The signed agreement to NTL is a legal document, and an online page will be developed to show the status of compliance with the data submission requirement for each industry platform deployment in the Gulf. NDBC will continue to provide the WMO ID and it will be used in all official labels. A monthly summary of data submissions will be reported to BSEE for their perusal and management. The report will include summary of data types and the number of valid and data rejected during the QC process.

GCOOS does not maintain records on when data are transmitted or received. However, the SFTP server log will be in position to record all connections made and this can be mapped with the registered IP (see Task 3.2) and used for the monthly reports. The server logs will be maintained and stored separately.

Task 6. Disaster and Recovery

This supplemental task to ongoing data collection activities of GCOOS will adhere to IOOS and GCOOS standards in the Disaster and Recovery (D/R) strategy.

Task 6.1. Near-site Backup Storage

A near-site backup data storage service will be installed to facilitate a daily backup of data and for rapid recovery when needed. A direct physical copy of the data submitted will be used and copied to a compressed-folder (projected 800GB/year storage space requirement). All raw, CSV converted and NetCDF files will be archived by GCOOS.

Task 6.2. Offsite Backup Storage

An offsite data storage service will be installed to mirror the primary server's entire services (SFTP and webpages). An automated Domain Roll-over function will be established to rollover to the offsite server that also functions as the alternate server to ensure the high-availability of services.

Task 6.3. Archiving to NCEI

All data in the NetCDF data format will be submitted to the National Centers for Environmental Information (NCEI) for the long-term archival of the data using the established parameters already agreed upon by GCOOS with NCEI.

Task 6.3.1. NetCDF quality check

All data will be subjected to QARTOD's QC requirements. When needed, additional QC procedures will be made to the data and a new flag will be added. As is the current procedure in GCOOS, all data will undergo an internal quality check before archiving is made (Task 6.3.2). The NetCDF file global variables will reflect the company that provided the data and GCOOS as the publisher.

Task 6.3.2. Updating the manifest

As per the agreement with NCEI, the manifest from where the data are listed for data harvest will be updated with the computed SHA384 for each file.