Data Management Plan

Last Updated August 28, 2024

Introduction

The Great Lakes Observing System (GLOS) is certified as one of eleven Regional Coastal Observing Systems (RCOS) for the U.S. Integrated Ocean Observing System (IOOS®). IOOS is governed by the Integrated Coastal and Ocean Observation System Act (ICOOS Act) of 2009 (33 U.S.C. 3601-3610) and the Coordinated Ocean Observation and Research Act (COORA) of 2020 (amending 33 U.S.C. 3601-3610) and operates under the lead of the National Oceanic and Atmospheric Administration (NOAA).

According to the ICOOS Act and COORA, the term "`regional coastal observing system' means an organizational body that is certified or established by contract or memorandum by the lead Federal agency designated in section 12304(c)(3) and coordinates Federal, State, local, tribal, and private interests at a regional level with the responsibility of engaging the private and public sectors in designing, operating, and improving regional coastal observing systems in order to ensure the provision of data and information that meet the needs of user groups from the respective regions."

GLOS, as the certified Great Lakes RCOS, works to aggregate and provide public access to Great Lakes observing data and information. GLOS data management system supports, among other things, data collection, aggregation, and quality-control from a variety of regional observational and numerical model assets. GLOS certification ensures that data management is done according to standards set by the federal government.

Members of the Data Management and Cyberinfrastructure (DMAC) team at GLOS work closely with the IOOS office and other RCOS to contribute to the development, testing, and implementation of standards as needed. GLOS follows all data management protocols promulgated by the IOOS office and works to implement them if needed, as soon as practicable limited only by resource restrictions, but well within a year of adoption by IOOS

GLOS, being a certified Regional Association member of IOOS, automatically ascribes to the GEOSS data sharing principles and commits to maintain data access following the **Find, Access, Interoperate, and Reuse (FAIR) data principles. GLOS also ascribes to the CARE principles (Collective benefit, Authority to control, Responsibility, and Ethics) for Indigenous Data Governance**.

GLOS DMAC Team

The GLOS Data Management and Cyberinfrastructure (DMAC) team is responsible for the requirements gathering, design, development, maintenance and support of GLOS information technology platform, Seagull, and other apps that serve the GLOS users.

The GLOS DMAC Team consists of three full-time staff, a full-time subcontracted Senior Advisor:

- The **Chief Information Officer** is responsible for leading the overall vision and strategy for the planning and execution of the DMAC infrastructure, software development via employees and contractors, related initiatives applications, technology platforms, and supporting technical development.
- The Senior Advisor, fulfilling the **'Data Services Manager'** role, is responsible for providing evaluation and coordination of the GLOS data management team of staff and contractors, manages GLOS data management services, serves as the GLOS representative for IOOS DMAC responsibilities, and assists in GLOS' strategy in developing the next generation IT platform.
- The **Cyberinfrastructure Engineer** is responsible for supporting the GLOS Cloud infrastructure, services, data streams, data storage and related functions.
- The **Marine Geospatial Analyst** serves as the central geospatial data expert at GLOS and is tasked with supporting the observation, data management and cyberinfrastructure, and communications teams.

The capacity for the GLOS DMAC team is a combination of in-house expertise and contract vendors providing high value development and operational support in a cost-effective manner. The subcontracted technology vendors are selected via a rigorous and open proposal and qualification process. This dynamic model allows for flexibility while providing maximum value to deliver cutting technology under strict oversight by the GLOS DMAC data management team.

GLOS DMAC Technology

Seagull

GLOS has developed Seagull, a next generation information technology platform, to serve data and information for a wide range of users, partners, stakeholders and constituents in and around the Great Lakes. Seagull is developed and deployed in a Cloud based environment (Amazon Web Services), leverages additional Cloud based services for the Seagull infrastructure and was released for public use in April, 2022.

Seagull is a free platform to use, available to the public and adheres to IOOS Data Management and Cyberinfrastructure core capability requirements including the FAIR principles (Find, Access, Interoperate and Reuse). Seagull is accessed via API's (Application Programming Interfaces),

Metadata discovery and catalog utilities, a web application and data, image, feature, basemap and other web services.

Seagull is an Internet of Things (IoT) marine information technology platform. It provides a broad set of users a discovery utility for Great Lakes data and information. It does this through the ingestion, management, dissemination and archiving of data generated from the observation network in the region. There are primarily 3 pillars to Seagull: 1) Apps & Services, 2) Enabling Technologies, and 3) Data. Interfaces serve as a 'connective tissue' facilitating the transfer of data, information, services and credentials between the pillars.

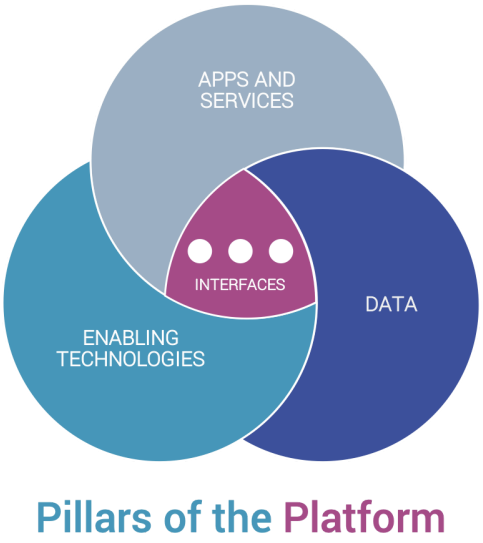

Figure 1: Pillars of Seagull

Interfaces

At the core of Seagull are Application Programming Interfaces (APIs). These interfaces consist of APIs and serve as the 'connective tissue' between the major pillars of the platform. Apps and Services, enabling technology, data, software upgrades and Cloud to machine communications pass via these interfaces. This critical component of the platform is scalable, secure and serves as a basis for integrating technology, data tunneling and communications from device to Cloud.

Enabling Technologies

Seagull serves as an environment where technologies can be integrated. GLOS DMAC architecture allows 3rd party developers and vendors to integrate additional functionality to the platform, extending and improving it.

Data

GLOS is a part of IOOS and therefore abides by certification rules and periodic audits. As a result, data storage, quality control, metadata attribution, and traceability are core values to the GLOS DMAC data strategy. Seagull uses multiple tiers of storage and retrieval based on business value, ranging from streaming services to near real-time access, access, archive, and deep archive.

Seagull also handles data structures that are both vector and raster spatial data. Accompanying metadata adheres to stated metadata standards and protocols (agency dependent). Volumes from datasets range from just a few bytes to many terabytes.

Apps and Services

There is a wide range of workflows and communities contributing to and consuming data and information from Seagull. These workflows range from registering smart devices and sensors to the receipt of notification-based alerts on smartphones. Seagull also allows third party organizations to tap into Seagull to augment their own application development with data and information services. “Branded” versions of Seagull allow organizations to leverage and re-use the back-end architecture and functionality of the GLOS technology platform to serve their own needs with some customization possibilities.

Seagull also supports the inputs and outputs required of internal and external application development, embedded services and the ability for other individuals and systems to consume data and information services.

GLOS aims to maintain an uptime of 99.9% for all the cloud infrastructure that runs these services. GLOS has set up internal dashboards and services that help monitor the applications, respond to system-wide performance changes, and optimize resource utilization on Seagull.

add language about distinction between operational center vs. regional association (ask sneha about her follow up with noaa) + we have no requirement as an IOOS CN to have required uptime

Data & Information Flow

Data Providers

Seagull ingests data from a number of regional data contributors:

- GLOS funds, fully or partially, the collection of a large amount of regional data, and these datasets must meet basic requirements for metadata, file format, and QA/QC.
    - Support for the management of observing assets, fully or partially funded by GLOS, is expected to be transitioned overtime to other partners. The expectation

is that the data collected with these assets will continue to be shared with GLOS to serve stakeholders needs as initially intended.

- GLOS can also ingest available, relevant data from federal, state, and provincial data providers, not funded by GLOS, and these agencies have required QA/QC protocols.
- GLOS can also ingest data from individuals, academic institutions, private, and non-profit organizations, not funded by GLOS.

Types of Data

Seagull ingests real-time, near real-time, and delayed mode or historic observing data, model data, and other data types including bathymetry and more.

Real-time data, provided mostly through buoys, moorings, and shore-based platforms, is normally transmitted on regular intervals to the GLOS streaming application programming interface (API) endpoints via HTTP(s) or through third party cloud systems.

Near real-time data normally requires some level of processing on the part of the data provider and that makes their availability delayed.

Delayed-mode or historical data, which are usually individual or aggregated datasets collected over time, are uploaded to Seagull in a variety of formats such as csv, xls, netcdf, pdf and more.

Forecast data from federal and/or GLOS funded partners are also ingested periodically from the partner systems such as TDS in netcdf formats and then processed into user readable formats such as geoJSON.

Proper data documentation or metadata is required for all data ingested by GLOS. The metadata requirements GLOS implements follow IOOS Metadata Profile Version 1.2. Among the metadata attributes required or recommended by GLOS are a description of the platform, dataset, and team that collects the data, as well as the dataset POC's email, institution, and url. GLOS adheres to the following vocabulary standards: the Climate and Forecast (CF) conventions, and IOOS Parameter Vocabulary 2.0.

Quality Control

The quality assurance procedures that GLOS and its partners implement help ensure that data delivered by GLOS is as accurate and precise as possible. GLOS funding agreements with regional partners require that they implement proper quality checks. In particular, observing equipment must be calibrated, operated, and maintained in accordance with manufacturer guidelines and recommendations, industry standards, and/or national IOOS program guidelines, when available and appropriate. Data services operated by federal agencies are assumed to

comply with the corresponding agency QA/QC policies. GLOS monitors data transmission to screen for transfer errors but does not perform quality checks on the data.

Regarding real-time data GLOS follows IOOS requirements to implement QA/QC checks according to a series of IOOS Quality Assurance for Real-Time Oceanographic Data (QARTOD) manuals. These manuals include a series of required, highly recommended, or recommended quality checks. GLOS runs, at a minimum, required QARTOD checks on all incoming real-time data for which protocols exist. The results from the quality checks are incorporated into the corresponding datasets and made accessible to users on the GLOS ERDDAP endpoint.

GLOS partners process delayed mode or historical data and run quality checks, which in many cases include many of the QARTOD tests. QA/QC information for these datasets, if made available to GLOS, is referenced in the associated metadata records.

Data Processing

The GLOS DMAC system manages and operates a collection of data processing pipelines that supports the variety of data types that GLOS system ingests. The pipelines developed are expandable and scalable to work with new data formats and integrations with other systems such as IOOS data assembly centers and NOAA cloud systems. Each data processing pipeline has well defined entry, integration, and exit points that are supported by numerous APIs that support both internal and external data needs of the system.

A typical data processing pipeline in the system ingests the data into the system, adds ISO 19115 compliant metadata to the incoming data, applies IOOS QARTOD quality checks wherever applicable, removes duplicate data, converts the data into canonical formats that can be stored in the database, and re-packages data to be made available to the DMAC system's peripheral applications such as the front end app, ERDDAP, and other API endpoints as applicable.

Data Storage and Archival

GLOS typically works with two types of data: data that are stored and hosted on GLOS cloud assets and the data that are only cached to provide access to the users through the front end application, avoiding redundancy of data storage. The former mostly includes real-time and near real-time observation data collected by the platforms deployed in the Great lakes region. The latter includes datasets that are already hosted on federated and non-private databases such as IOOS DACs, National Data Buoy Center, NOAA cloud databases, and others.

The datasets that are hosted by GLOS are stored on the cloud in databases that are backed up periodically. Furthermore, as an IOOS certified region, GLOS also archives all the real-time datasets at the NOAA National Centers for Environmental Information (NCEI) database monthly. The observational data for each platform are aggregated and converted into NetCDF files which

are then pulled by NCEI and archived on their systems. This ensures long term archival of the real-time observation data that are hosted at GLOS systems.

Data Dissemination

Part of GLOS' mandate is to make Great Lakes data findable and accessible to the public. This is achieved through comprehensive metadata records, a user-focused information platform, and data integration with federal platforms wherever possible. Data and metadata can be accessed primarily through our front end application that includes all types of data collected and aggregated by GLOS. The front end application also references our metadata database, which is an Esri Geoportal Server Catalog instance supported by Amazon Web Services (AWS) OpenSearch, that hosts all of the metadata records collected by GLOS. This also allows for discoverability of the datasets that may not be hosted at GLOS but are accessible through other portals and apps. GLOS also has other open API endpoints that allow users to query the datasets as specified in the open specifications.

GLOS also maintains a standard service called ERDDAP that allows users to programmatically access the datasets through a catalog of RESTful application service endpoints. GLOS' ERDDAP is a highly functional data access application that also works as an integration endpoint between GLOS and national data centers or catalogs. In particular, all the observation data that are ingested by GLOS is served on GLOS ERDDAP service endpoint which are then integrated into the IOOS Catalog and are made available on the IOOS sensor map for nationwide access.

Also, in partnership with IOOS, NOAA NDBC ingests nonfederal IOOS partner data and delivers a subset of those variables (mainly meteorological and physical oceanographic) through the National Weather Service system and onto the World Meteorological Organization (WMO) Global Telecommunication System (GTS). All data published to the GTS are accessible to any climate and weather forecast center. NDBC also publishes all the data they harvest to their web products.

As a service to the data contributors, GLOS provides integration between the GLOS system and NDBC to allow NDBC to pull data in real time from the platforms that have a valid WMO ID. NDBC has recently moved to leverage IOOS RA ERDDAP servers, in this case the GLOS ERDDAP server, as the central access point for nonfederal IOOS partner data. This is a departure from the previous data delivery method, where providers packaged XML files and posted them to an NDBC FTP server. Refer here for more details. It is critical that the data contributors work with GLOS to ensure their data are available in the GLOS ERDDAP service as required by NDBC, so that NDBC can continue to access those datasets.

Any data received by GLOS with stipulations that data not be made public, will not be made public via GLOS products **or** shared with the federal government, unless GLOS is notified by the data provider of their desire to do so.

Data Security

GLOS encourages all the data contributors to transfer data, or otherwise ensure all data contributed are using protocols with the best possible security. Methods range from the use of HTTPS to use of authentication keys granted to providers prior to onboarding of their platforms. Furthermore, necessary authentication and per-user privilege mechanisms are implemented to make the data secure at all times within the architecture landscape without compromising on the open data sharing policies of GLOS.