**Data Management Plan NES-LTER Phase II**

This data management plan is organized first to summarize how the NES LTER information management (IM) team supports the full data lifecycle and then to step through the DMPTool template provided by NSF's Biological and Chemical Oceanography Data Management Office (BCO-DMO). Our IM team includes a lead Information Manager, a technical staff member who serves as 'at-sea' data lead, a technical staff member to facilitate hydrographic data processing, and staff in WHOI's Information Services (IS; as described in the Facilities Statement).

**Infrastructure to support the full data lifecycle**

A primary goal of NES LTER IM is to support the project's research activities by facilitating the data lifecycle (Fig. DMP 1).

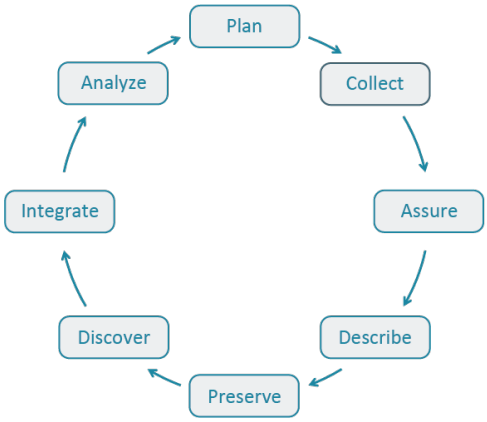


**Fig. DMP 1.** DataONE Research Data Lifecycle Diagram.

- Plan - Our IM team participates in science planning meetings, coordinates data acquisition prior to and following cruises, and trains project teammates in best practices for data management.
- Collect - Our project collects "Big Data" in terms of *volume, variety, and velocity,* requiring a complex infrastructure described below.
- Assure - To address *veracity* our IM team employs quality assurance in our data product workflows including IODE (International Oceanographic Data and Information Exchange) quality flags for some products.
- Describe, Preserve, Discover - To contribute FAIR (Findable, Accessible, Interoperable, Reusable) data products to DataONE and other community repositories, our IM team uses non-proprietary data formats, standardizes metadata, and promotes the use of controlled vocabularies.
- Integrate and Analyze - Our web-based REST API (representational state transfer application programming interface) for cruise data and our sharing code online in GitHub allow our project team to quickly build workflows for data integration, analysis, and visualization.

Our infrastructure must balance the project participants' need to have efficient access to large volumes of data with the additional goal of providing public access to data products. Our solution involves components spread between on-premises and cloud providers (Fig. DMP 2, Table DMP 1). By mid-year 6 of our project, our IM team stores ~60 TB of data (~55 on premises and ~5 cloud). For WHOI's institutional research data storage (RDS) on premises, we have >40 TB to share within WHOI a subset of our towed plankton imaging data, 12 TB for a subset of our coupled physical-biological model output, and 1 TB for (low-volume) data products including those served by our web-based REST API. However, additional storage is required for plankton imagery (e.g., PI Sosik's network attached storage serving ~10 TB publicly available IFCB data), acoustic data collected by vessels and Stingray towed vehicle, and physical and coupled biological-physical model output; these and some other high-volume data types (e.g., high throughput sequencing data) are necessarily managed by PIs. Much of the high-volume data

needs to be proximate to high-performance computing resources. Cloud storage is used for Google Drive, GitHub, and Zotero (in decreasing order of total data volume). Google Drive is our primary means of sharing files among project participants across institutions and is used to store some raw data and internal documentation. GitHub is used to store IM-related software including data package assembly for community repositories, and we maintain a list of project participants also using GitHub to share code. Zotero is used to organize our project's products including published articles, data, and conference presentations for citation.

Many of the IM-managed components are intended to effectively handle and distribute the data collected during transect cruises (and post-cruise from samples collected). Once we obtain the external drives from these cruises, the IM copies ship- and PI-provided data into Google Drive and conducts initial quality assurance prior to uploading a subset of data types into WHOI's RDS for scientists and students to be able to use our web-based REST API to integrate cleaned data into their data visualization and analysis workflows. The API provides the ability to read data directly into code and is language agnostic to accommodate the use of a variety of programming languages (Matlab, Python, R). The IM team uses the API when compiling core long-term data products for curation and submitting to community repositories. IM-managed data products for our REST API and the project website are housed on a suite of virtual servers maintained by WHOI IS.

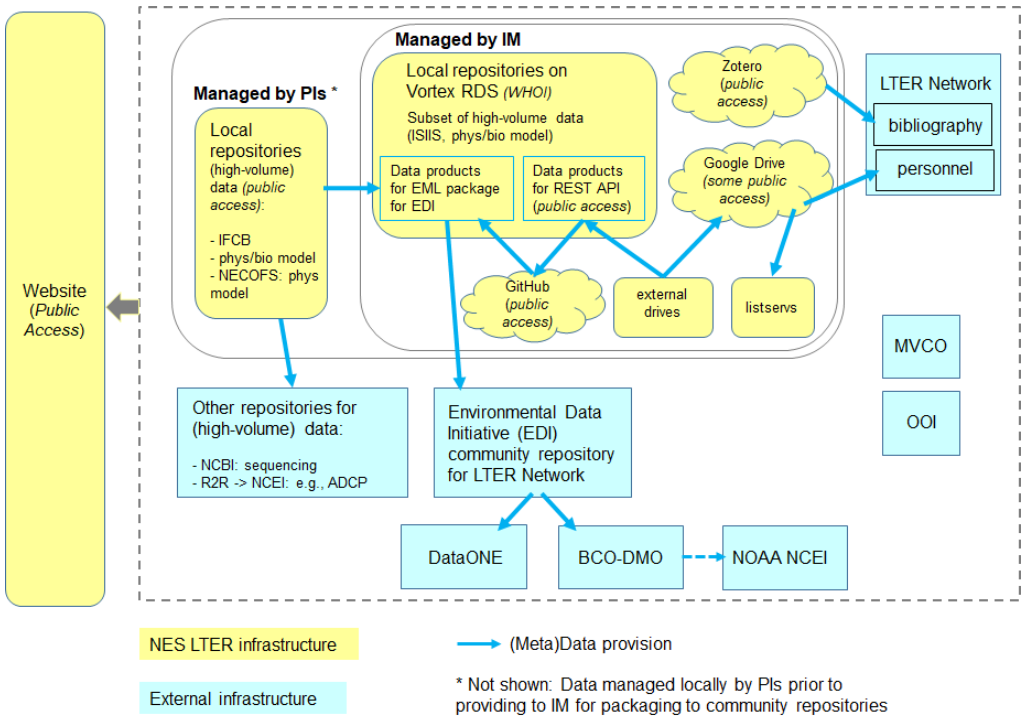


**Fig. DMP 2.** Major NES LTER information management system (IMS) features, with arrows indicating the flow of (meta)data to repositories for public access. Not shown: Provision of metadata from external repositories to our Zotero data catalog; Trello (cloud service) used by IM to document data packaging.

**Table DMP 1.** Major NES LTER information management system (IMS) features. [1]Feature required by all LTER sites; [2]further described in text; [3]provided by WHOI Information Services. Note: most of our data products served publicly through NES LTER REST API are incorporated into long-term core data products published to community repositories.

| Type | Feature | Implementation |
|---|---|---|
| Website, catalogs, and/or directories | [1] https://NES LTER.whoi.edu/;<br>[1] Bibliography and [1] Data catalog;<br>[1] Personnel directory and listservs;<br><br>[1] Protocols | [3] WordPress;<br>Zotero;<br>Google sheet > csv > LTER Network and 6 listservs;<br>Google Drive |
| Data products published to DataONE community repositories | [1] 52 data products publicly available through DataONE repositories (as of February 2023) | EDI; R2R; NCEI; Dryad; data in EDI and R2R linked to BCO-DMO |
| Data products served by other repositories | 15 data products publicly available through other community, institution, and local repositories | NCBI; SeaBASS; [2] PI-managed high-volume datasets with public access |
| Servers and user accounts | Cloud file system across institutions for project team; On-premises institution file system for sharing within WHOI | [3] Google Drive; [3] Vortex RDS |
| Code repository, Issue tracking | Code storage with version control for REST API and EML package production | GitHub; Trello |

**DMPTool template from BCO-DMO**

**Data Policy Compliance**

Our project will comply with the data management policies described in the NSF PAPPG (NSF 23-1) Chapter XI.D.4 “Dissemination and Sharing of Research Results,” the NSF Division of Ocean Sciences (OCE) Sample and Data Policy, and the Long Term Ecological Research (LTER) Network Data Access Policy.

**Pre-Cruise Planning**

Our project involves research cruises on vessels within and external to the University-National Oceanographic Laboratory System (UNOLS) fleet. For UNOLS cruises: pre-cruise planning will be done via teleconferencing. Detailed plans for station locations, instrument deployment, and water sampling strategy will be written up as a cruise plan. The actual sampling events will be recorded in a digital event log using the R2R event logger application, and any paper logsheets will be scanned into PDF documents.

**Description of Data Types**

We will produce observational data, experimental data, derived data products, and model data. Observational data will be obtained in near-real-time from moored underwater instruments, underway and from sampling on research cruises, and post-cruise with laboratory analyses of physical samples including water samples, filters, plankton net samples, and fish specimens. Experimental data will be generated from incubation experiments conducted during research cruises. Derived data products will include rates calculated from observational and/or experimental data. Model data and post-processed model data products will result from coupled biophysical modeling. Data products will be categorized into the 5 LTER core areas, as in the table of datasets from NES-LTER Phase I (supplementary document).

**Data and Metadata Formats and Standards**

Data products provided to DataONE repositories will be in non-proprietary formats when applicable, e.g., comma separated values (CSV); "raw" data in formats as collected will also be provided when applicable. Metadata will be provided to the EDI repository in the Ecological Metadata Language (EML) standard. Our primary means of curating data with EML metadata uses the emlassemblyline R package for "ongoing" data packages, with utility functions and a spreadsheet template to streamline the incorporation of metadata contributed by lab groups; for "completed" data packages, we use the ezEML tool. We strive to meet FAIR data standards with effort towards high quality metadata and method details for each dataset. Other data and metadata standards may apply to some high-volume and/or high-frequency data. For example, high-throughput sequencing data will be formatted for the National Center for Biotechnology Information (NCBI), and model data will utilize data and metadata formats familiar to the U.S. Integrated Ocean Observing System (IOOS) community (e.g., NetCDF format). We select attribute names from controlled vocabularies including those served by the British Oceanographic Data Centre (NERC Vocabulary Server). We employ IODE (International Oceanographic Data and Information Exchange) primary quality flags. We select dataset keywords from controlled vocabularies including the LTER Controlled Vocabulary.

**Data Storage and Access During the Project**

As detailed in the section above ("Infrastructure to support the full data lifecycle"), data storage and access during the project differ depending on data type and volume. In general we strive to use resources that enable sharing data across institutions - thus, our use of Google Drive and other cloud resources, and our REST API for a subset of transect cruise data in WHOI's Research Data Storage (RDS), as shown in Fig. DMP 2. Some of our high-volume data types are locally served with public endpoints (e.g., IFCB data); however, other high-volume data types are stored in infrastructure local to respective institutions, also shown in Fig. DMP 2. The IM will ensure regular backup of shared cloud resources (e.g., to external drive); backup for the IM-managed storage on WHOI's RDS is provided through Commvault to either tape storage or Amazon Deep Glacier in the cloud. PIs are responsible for backup as described in each institution's Facilities Statement.

Our full data catalog is implemented in Zotero, linked from our project website's Data page, where we also provide a suggested order in which to find and access NES LTER data (in decreasing order of curation): (1) Check to see if data are curated and published at EDI (or another community repository); if not yet, then (2) check to see if data are available through our REST API; and if not in that subset of data products, then (3) for project participants, access data in our project Google Drive.

**Mechanisms and Policies for Access, Sharing, Re-use and Re-distribution**

With regard to public access to (meta)data and other relevant digital products, we aim to publish our data products through DataONE community repositories, with some products (e.g., high-volume data from images or models) served by other repositories (Table DMP 1). We provide long-term datasets to the LTER Network's repository—the Environmental Data Initiative (EDI) repository—with some exceptions, e.g., if another community repository in the ocean sciences is more often used for a particular type of data. NES-LTER data will be made freely and publicly available following guidelines from the LTER Network Data Access Policy for Type I data. To meet the time frame of release within 2 years from collection, the IM team maintains a data publishing 'priorities' sheet and has regular meetings to iterate ongoing packages and document completed packages. We also have engaged software engineers to assist with automation and to engage in technical discussions (e.g., with EDI) to develop streamlined ways of dealing with data variety and volume. Some high-volume and/or high-frequency data will be provided through other community, institutional, or local repositories. For example, high-throughput sequencing data will be provided to NCBI; underway data from UNOLS cruises will be provided through the Rolling Deck to Repository (R2R); underway data from Tioga coastal vessel day cruises will be provided through WHOI's institutional repository. Our project landing page at NSF OCE's BCO-DMO points to our data packages in EDI as well as our cruise data in R2R. With regard to licensing data for re-use, we recommend Creative Commons CC0 – No Rights Reserved or CC BY – Attribution when possible as recommended in the LTER Network Data Access Policy. Type II data restrictions might apply to products from remote-sensing data if covered under prior licensing.

**Plans for Archiving**

We will facilitate the archiving of data with NOAA NCEI when possible, and with WHOI's Data Library and Archives when applicable. For those data provided to other community repositories, archiving plans of those repositories apply. Local repositories at WHOI's data center will include backup and disaster recovery as described in the Facilities Statement. PIs will archive voucher specimens in their labs.

**Roles and Responsibilities**

Each PI will be responsible for sharing his/her subset of data among the project participants in a timely fashion. The lead Information Manager, supervised by Lead PI Sosik, will coordinate the submission of ongoing (long-term) data products to the EDI repository and will facilitate and document other data products submitted by project participants to appropriate repositories.