Data Dissemination and Cyberinfrastructure Ad Hoc Committee Initial Report

29 September 2017

1 Introduction

To facilitate the assessment of Ocean Observatories Initiative (OOI) data quality by the scientific community, and to accelerate the integration of OOI infrastructure usage into project proposals and scientific publications, the NSF Ocean Observatories Initiative Facility Board (OOIFB) established the Data Dissemination and Cyberinfrastructure (DDCI) ad hoc committee which is tasked with identifying near-term and longer-term obstacles to the enhanced delivery of data to the scientific community and providing recommendations for removing these obstacles.

The DDCI comprises the following individuals:

- Timothy Crone, LDEO (co-Chair)
- James O'Donnell, UConn (co-Chair)
- Brian Glazer, UH
- Orest Kawka, UW
- Stephanie Petillo, WHOI
- Mary Jo Richardson, TAMU
- Richard Signell, USGS
- Derrick Snowden, NOAA IOOS
- Larry Atkinson, OOIFB Chair (ex officio)

The DDCI met in-person at NSF headquarters on 18 July 2017, and has had several Webex/conference calls prior to and following this meeting. NSF program manager Lisa Clough was present for the meeting and all calls. Other representatives from NSF including Rick Murray, Bob Houtman, and Rachel Shackleford were present for part or all of the in-person meeting. Annette DeSilva (UNOLS/URI) facilitated meetings and calls and took meeting notes.

During the meeting and the calls, the committee spent a significant amount of time learning about the current state of the OOI cyberinfrastructure and how data is currently handled and disseminated. We heard presentations from Mike Vardaro of the Rutgers Data Team and Ivan Rodero of the Rutgers CI Team. We heard a presentation by Rich Signell on ERDDAP, a presentation by Stephanie Petillo on OMS++, an alert and alarm system for OOI that uses ERDDAP as a data back-end, and a presentation by Tim Crone on the OOI high-definition camera system (CAMHD). We had lively and informative discussions about the needs of the scientific community, potential new modes of data access for the scientific user, and discussions regarding potential improvements to the management structure of the CI.

This report is a summary of the committee's findings representing our views at this stage of our efforts. The committee expects to continue our work on this problem, to meet again in the future, and to refine our views and recommendations as we learn more about the current state of the system and receive input from operators and the scientific community.

The findings and recommendations in this report are broken down into two sections. In the first section, we detail our recommendations for short-term adjustments to the current cyberinfrastructure priorities and management structure that we believe can be reasonably accomplished in the next few months. In the second section, we detail our recommendations for "OOI 2.0" which include longer-term recommendations that should be considered as the next phase of OOI operations is planned and a new Cooperative Agreement (CA) for the management and operation of OOI is formed and executed.

2 Near-Term Recommendations

The committee has several near-term recommendations to facilitate data dissemination in the coming months, which the committee thinks can be reasonably accomplished before the transition to OOI 2.0. These recommendations are:

1. Prioritize the development and public release of the uFrame-powered ERDDAP server.
2. Accelerate the ingestion of backlogged data.
3. Identify a single individual who reports directly to the Project's Lead PI, who will be responsible for data access by scientists and who has authority over both CI and Data Team priorities.
4. When the ERDDAP system is fully functional, documentation for the CI system must be completed and all code made available in publicly-accessible repositories.

2.1 ERDDAP

ERDDAP is a free and open-source Java "servlet" which for the non-expert can be thought of as a kind of specialized web server that excels in converting and serving disparate scientific datasets using a uniform interface. ERDDAP is focused primarily on serving gridded or tabular (e.g. time-series, profile) datasets which are commonly stored on the server as NetCDF files, and it can serve data in a large number of formats as well as generate plots and maps of requested data. Together, ERDDAP and NetCDF allow data, metadata, and data attribution information to be distributed in a convenient and efficient manner.

ERDDAP has a standard browser interface that facilitates searching for, converting, and plotting data, but ERDDAP is built on a RESTful machine-to-machine API, meaning that the server does not store browser state and all information about every request is contained in the URL of each request. This makes it easy to automate searching for and using data in other applications like Python, R, JavaScript, or MATLAB, and makes it easy for users to build their own custom interfaces if they so wish.

ERDDAP is a server framework that allows anyone with data to serve their data by running their own ERDDAP server. Many dozens of organizations (including NOAA, NASA, and USGS) are now running ERDDAP servers to serve their scientific data, and ERDDAP is on its way to becoming a de facto standard in the Oceanographic community. The ERDDAP principal developer and user community have created user guides, instruction videos, and code examples to facilitate access by new users.

The committee believes that ERDDAP has the potential to serve most of the OOI data in an efficient and useful manner and that the deployment of an ERDDAP system that works on top of uFrame could greatly expand OOI data availability for the scientific community. **The committee recommends that the development of the ERDDAP system be made the top priority for the near term.**

To expedite the development of the ERDDAP system, the committee recommends that the ERDDAP development team be provided with the access they need to complete this task as quickly and as efficiently as possible. At a minimum, the ERDDAP team should be given read access to the production log files. Another suggestion for speeding up the development of ERDDAP is to reduce the deployment timeline from two weeks to a few days, specifically in support of the ERDDAP team to accelerate the development of this system.

The committee also recommends that the ERDDAP developers begin or continue to interact with other OOI developers such as the CGSN who have developed internal ERDDAP systems, and that they ensure that the ERDDAP data sets are well-described using best practices for international standards. For example, it would be best if the OOI CI way of publishing data followed the NCEI NetCDF Templates, widely used in the community. These were designed for long-term preservation, scientific quality control, product development, and multiple data re-use beyond its original intent.

2.2 Data ingestion

Ingestion backlogs are an area of concern in terms of data availability for the scientific community. **The committee recommends that data ingestion remain a top priority for the CI and Data Teams.** The committee notes that although the M2M ingestion system appears to be promising, the MIOs are not currently using it, and may in fact not be authorized to use it. Also, it is not clear how a distributed ingestion model can work. The committee recommends that the CI and Data teams continue to focus on data ingestion using a centralized model with the understanding that the MIOs may not be involved in the near-term.

2.3 Data Delivery Manager

It is the committee's view that the organizational structure of the CI and Data team has created roadblocks to the effective and efficient dissemination of data to the scientific community, and inefficient allocation of resources. Further, collaboration with the MIO personnel appears to have been hindered since useful tools and experiences have often not been effectively shared. **Since the primary motivation of the OOI is to deliver data to scientists, the committee believes that the program would benefit from the establishment of an OOI Data Delivery Manager.** The Data Delivery Manager should report directly to the Lead PI and be responsible for the primary product of the OOI. To be effective, the Data Delivery Manager must have authority over the work that is currently conducted by the CI and Data Teams, and also have frequent interaction with the technical staff at the MIOs. As data reaches more scientists, issues that require technical input from the entire OOI will emerge and responses coordinated. For the OOI to be successful, the Data Delivery Manager must have the resources and authority to ensure that the system is responsive to the users' needs and input.

2.4 Document the cyberinfrastructure

Although there is a large amount of existing documentation regarding the data, it seems that many of the essential tasks of running the existing cyberinfrastructure are known only by specific key personnel. Once the ERDDAP capability has been added to the system, the focus of the CI team should shift to completing the documentation that will allow development, maintenance and knowledge transfer to other developers and system operators. The documentation needs to cover all aspects of the cyberinfrastructure, including server environment, installation and configuration, data workflow components, troubleshooting, and code development. Further, barring well-justified reasoning on a repository-by-repository basis, OOI should begin the process of moving code back into publicly available GitHub repositories.

3 Long-Term Recommendations (OOI 2.0)

The committee has several longer-term recommendations to facilitate data dissemination as OOI 1.0 transitions to OOI 2.0. These recommendations are:

1. Assess the future viability of uFrame.
2. Place a primary focus on satisfying the data needs of the scientific user base.
3. Consider partnerships for providing remote compute capability for larger OOI datasets.
4. Maintain a Data Delivery Manager in OOI 2.0.
5. Support operational centers by disseminating data in real-time via the Global Telecommunications System and other systems used by the operational community.

3.1 Assess the future viability of uFrame

The committee and nearly everyone consulted by members of the committee have serious concerns about the uFrame system. One primary concern is that uFrame in effect places a "black box" or at best a "gray box" in the processing pipeline, and it is difficult for end users to fully ascertain how data products are

generated from raw data. It is difficult if not impossible for users to run custom processors using different calibrations or other processing parameters to experiment and troubleshoot. This lack of obvious transparency has caused many members of the scientific community to express healthy levels of skepticism regarding the data pipeline.

Another primary concern is the apparent lack of documentation for uFrame and the proprietary nature of many components of the uFrame codebase. In addition to the obvious transparency issues when dealing with closed-source code, the proprietary aspect of the software may become a budget issue in the future. If there are no funds for planned product improvements or if Raytheon is unwilling to make uFrame open source, then OOI could be locked in with a Raytheon product for the foreseeable future. Even if Raytheon does release the source code, there is no guarantee that the current CI team (or the new team if that changes) will have the skills to maintain and modify what would become a fork of the Raytheon product into the public domain.

Despite these concerns, some committee members (but not all) believe that that the uFrame/Cassandra database model offers some advantages that *may* not be easily replicated using a simpler file based system. The first among these is that uFrame stores instrument raw data in the database and delivers processed and higher-level derived data products on-demand (i.e., applies scalable processors to the data upon data request). This model would, in principle, allow users to apply custom processor files to the raw data to generate alternative data products during queries, however it is not clear if this capability has been realized. Currently changing processors appears to be a long and complex process which regular users do not have easy access to. Another advantage is that the current system is capable of ingesting, processing, and serving data from the Cabled Array in real-time, which provides substantial scientific value.

For OOI 2.0, the committee recommends that uFrame be evaluated in terms of the issues listed above, and that potential alternatives be considered. Any replacement systems considered should not descope the capabilities of the CI, at least not without consulting the scientific user base, and specifically should maintain and preferably extend the “compute on demand” aspect of the system, and should maintain the real-time ingestion/processing/service capability of Cabled Array data. In the committee's view, any new system should favor simple, modular and reproducible components with demonstrated community use over complex monolithic solutions. The committee notes that many components of uFrame are open-source, including Apache Cassandra, so one solution could involve replacing the closed parts of the system in favor of simpler open workflows. The committee also notes that other real-time and archive web services exist for the data types collected by the Cabled Array data and are commonly used in the community, for example Antelope, SeedLink and FDSN. There are also potential partner organizations that could provide these services (e.g. IRIS, NCEI) if viewed beneficial or cost-effective to OOI.

3.2 Place a primary focus on the scientific user base

The OOI has enormous potential for outreach and education, for use by the general public, the media, and by students of all ages. However, the viability of the observing system during these early years of operation will be dependent on proposal pressure from scientists in the community to use and expand OOI assets, and on the publication of peer-reviewed journal articles based on OOI data. Indeed, **a primary goal of this committee is to accelerate the availability of OOI data for scientists and thus expand the use of these data for science. For this reason, the committee recommends that efforts to improve the user experience on the OOI data portal, and the expanded availability of data through systems such as ERDDAP or the M2M interface be focused on the needs of working scientists.** Based on our discussions, our view is that scientists have needs and requirements that are quite different than the casual user, and can be summarized by this list of questions a scientist is likely to ask when looking to obtain data:

1. What data are available? What instruments are working and which ones are not? Scientists need an easy-to-see overview of the entire system to help them plan research activities.
2. How good are the data? Are the metadata flags easy to understand and are they well incorporated into the data provided? Scientists need to know how reliable the data they obtain are.
3. Where are the data? Are the data easy to download in easy to use formats? Can the data be downloaded by clicking a link instead of by waiting for an e-mail to arrive? Scientists may want to see plots of data in real-time, but in most cases scientists will want to download data in some sort of tabular format (e.g., HDF5) that allows them to do their own processing and visualization using the software tools of their choice.
4. Can the workflow from raw to processed data be reproduced easily and independently, so that new algorithms and approaches may be tested and improvements fed back into the processing system?

Efforts to improve the UI of the OOI should be focused on how working scientists actually use data and processes must be developed to solicit and consider user input, and evaluate effectiveness.

3.3 Consider partnerships for providing remote compute capability for larger OOI datasets

For some OOI data the download model for data access is simply not viable. In particular, the hydrophone and the HD video datasets are so large that researchers cannot hope to download these datasets within any reasonable timeframe. For example, the HD video dataset currently includes nearly 7000 high-resolution video files totaling approximately 85 TB in size. Not only might it take many weeks or months to download the entire dataset, but most researchers would struggle to find the space to store such a large amount of data locally.

For this reason, the committee recommends that collaborations and or partnerships be sought to provide combined compute and storage capability for these large datasets. One potential partner is XSEDE, a consortium of some of the country's largest supercomputer operators. XSEDE may be able to provide hosting and access to these data using some novel funding model. Other possibilities include the development of commercial partners such as Amazon or Google who may be willing to host these large datasets at affordable rates, or Calit2 which has expressed interest in hosting OOI data. Axiom Data Science currently houses large amounts of data for IOOS and other customers on their private cloud, and provides HPC capabilities within their Research Workspace, allowing scientists to process large datasets including hydrophone data in a scalable, data-proximate manner. The possibilities for partnerships abound, and for some of the data in the OOI system, a cloud-based solution is the best way to accelerate data access for the scientific community.

3.4 Maintain a Data Delivery Manager in OOI 2.0

A person should be identified that has the primary goal of overseeing data delivery to the scientific community, is responsive to the needs of the scientific community, and has oversight authority over all management components of the cyberinfrastructure system and administration so that decisions about the cyberinfrastructure can be made with the needs of the scientific community at the forefront.

3.5 Support operational centers by disseminating data in real-time via systems such as the Global Telecommunications System.

The Global Telecommunications System (GTS), run by World Meteorological Organization (WMO), provides a mechanism for the National Meteorological Centers to ingest and disseminate real-time observations and forecast products. Data collected by the OOI can be of high value to operational met services for data assimilation in real-time or model verification in delayed mode. GTS distribution from OOI can be facilitated by the NOAA National Data Buoy Center and the U.S. IOOS program. Some OOI glider data has already been submitted to the GTS through the IOOS Glider Data Assembly Center which

has been working with OOI CI Data Management teams on glider data submission recently. IOOS and NDBC are currently re-engineering the real-time data submission process to rely heavily on ERDDAP which is consistent with the recommendations in this report for the OOI-CI data dissemination strategy. Because of the alignment of technologies and given the high value of the OOI data to the operational centers, we recommend a firm commitment to distributing as much OOI data as is possible in real-time via the GTS as part of OOI 2.0.

4 Summary

In summary, the recommendations of the DDCI Committee over the short term are:

1. Prioritize the release of the OOI ERDDAP server by empowering the ERDDAP development team with all needed access to the production server and by shortening the deployment cycle timeline.
2. Accelerate the ingestion of backlogged data.
3. Identify a single individual, the OOI Data Delivery Manager, who reports directly to the Project's Lead PI and who will be responsible for data access by scientists and who has authority over both CI and Data Team priorities.
4. Make sure documentation is complete and all code is in publicly-accessible repositories.

To help guide the formation of the CA for OOI 2.0, the recommendations of the DDCI Committee are:

1. Assess the future viability of uFrame.
2. Place a primary focus on satisfying the data needs of the scientific user base.
3. Consider partnerships for providing remote compute capability for larger OOI datasets.
4. Maintain the position of OOI Data Delivery Manager in OOI 2.0.
5. Support operational centers by disseminating data in real-time via the Global Telecommunications System and other systems used by the operational community.