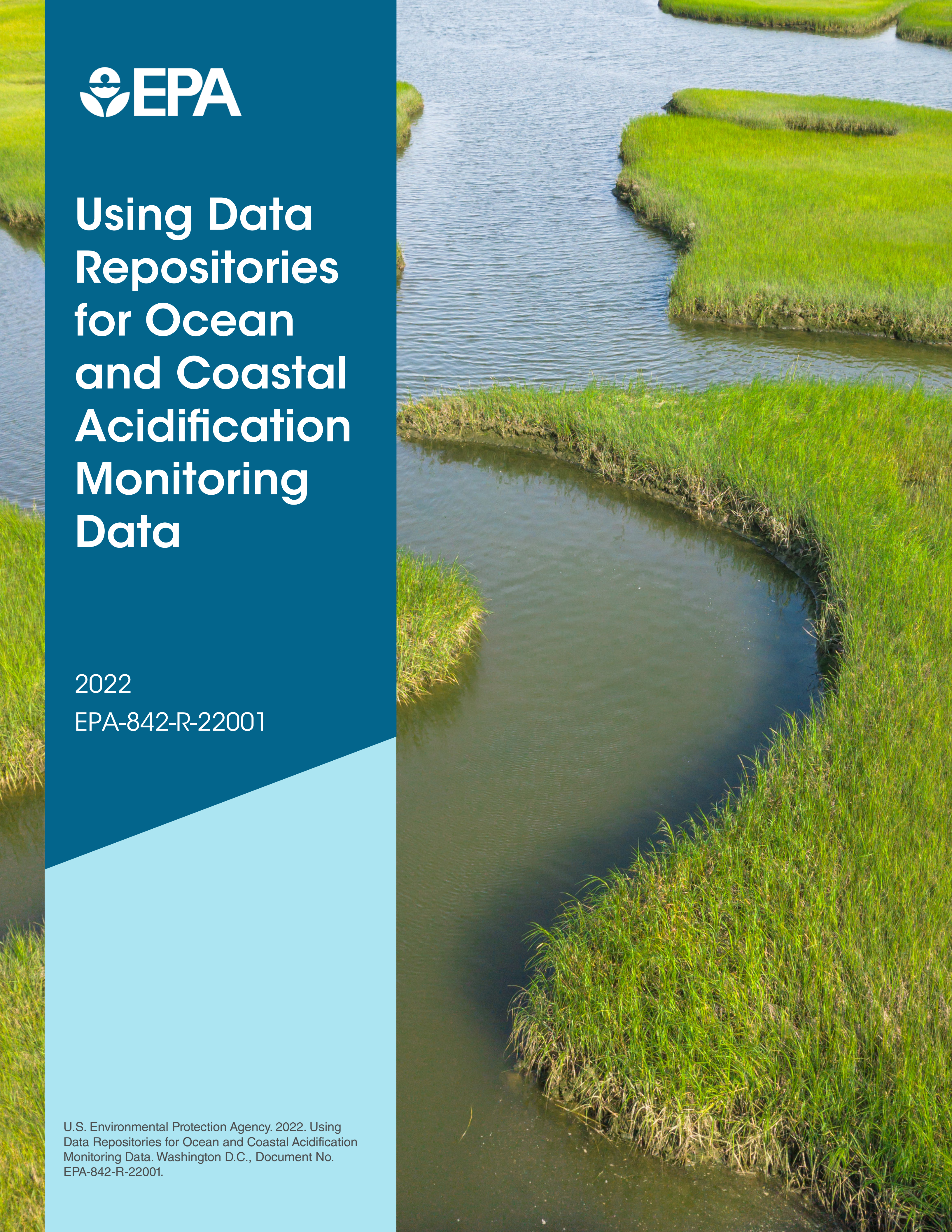


Using Data Repositories for Ocean and Coastal Acidification Monitoring Data

2022

EPA-842-R-22001

U.S. Environmental Protection Agency. 2022. Using Data Repositories for Ocean and Coastal Acidification Monitoring Data. Washington D.C., Document No. EPA-842-R-22001.

Authors:

Gulf of Maine Research Institute
Riley Young Morse

Northeast Regional Association of Coastal Ocean Observing Systems
Tom Shyka

EPA Office of Wetlands, Oceans, and Watersheds
Holly Galavotti

EPA Region 1 (retired)
Matthew Liebman

Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or EPA. EPA does not endorse any commercial products, services, or enterprises.

The views expressed in this report are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Contents

Acknowledgments

This project was supported by the U.S. Environmental Protection Agency (EPA) and developed in collaboration with the Casco Bay Estuary Partnership, Northeastern Regional Association of Coastal Ocean Observing Systems (NERACOOS), Gulf of Maine Research Institute, the participating National Estuary Programs and their partners.

Special thanks to the following individuals who participated in interviews about managing and accessing ocean and coastal acidification data:

- Dr. Curtis Bohlen, Director, Casco Bay Estuary Partnership
- Pam DiBona, Executive Director, Massachusetts Bay National Estuary Partnership
- Mike Doan, Research Associate, Friends of Casco Bay
- Dr. Chris Hunt, Research Scientist, University of New Hampshire
- Dr. Emilio Mayorga, Senior Oceanographer, Applied Physics Laboratory, University of Washington
- Dr. Karina Nielsen, Executive Director, Estuary & Ocean Science Center, San Francisco State University
- Nicole Petersen, Water Quality Specialist, Barnegat Bay Partnership
- Dr. Grace Saba, Assistant Professor, Rutgers University
- Dr. Jim Vasslides, Senior Program Scientist, Barnegat Bay Partnership
- Prassede Vella, Staff Scientist, Massachusetts Bay National Estuary Partnership

Test Case Participants:

- Mike Doan, Research Associate, Friends of Casco Bay
- Dr. Chris Hunt, Research Scientist, University of New Hampshire
- Jennie Rheuban, Research Specialist, Marine Chemistry & Geochemistry, Woods Hole Oceanographic Institution

Reviewers:

- Jonathan Pollak, Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI)
- Dr. Liqing Jiang, NOAA's National Centers for Environmental Information (NCEI) Ocean Carbon and Acidification Data Portal (OCADS)
- Katherine Weiler, EPA, Office of Wetlands, Oceans, and Watersheds
- Dwane Young, EPA, Office of Wetlands, Oceans, and Watersheds

Executive Summary

Since 2015, the U.S. Environmental Protection Agency (EPA) has funded and supported efforts to expand ocean and coastal acidification (OCA) monitoring to include *in situ*, autonomous pH and pCO_2 sensors at several National Estuary Program (NEP) sites. These NEPs and their partners have generated a large volume of high temporal resolution carbonate chemistry data and associated parameters including temperature, salinity, and dissolved oxygen. NEPs, however, often lack information, staff, and funding resources to store and share data beyond a local computer network. This results in limited discoverability and accessibility of data to a broader community of researchers and partners. Typically, data are made available through downloads from organization websites or by responding to direct requests for data. Continuous monitoring adds a new challenge where high frequency data result in large files, which cannot be readily transmitted and require greater storage capacity. Therefore, EPA commissioned this report to provide information that NEPs, their partners, and other monitoring groups can use to submit time-series ocean and coastal acidification data to publicly accessible online data repositories.

Several online data repositories were evaluated for submitting ocean and coastal acidification data collected through the NEP and other monitoring programs. The evaluation of the repositories was informed by interviews with NEP data providers and OCA community data users to better understand their data management needs. The findings from the interviews were synthesized to produce a set of attributes to evaluate the data repositories. Test cases were conducted with three data providers to evaluate their experience using data repositories. The evaluation identified NOAA's National Centers for Environmental Information (NCEI) Ocean Carbon and Acidification Data Portal (OCADS) and The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI)'s HydroShare Data Portal as the two most suitable repositories for ocean and coastal time series data because they are user-friendly and employ a robust approach to metadata requirements to make datasets discoverable, accessible, and interoperable. Other repositories that were evaluated include CUAHSI Hydrologic Information System (HIS), EPA's The Exchange Network (EN), and NOAA's Integrated Ocean Observing System (IOOS) Regional Association partnerships. It is important to note that this is not a comprehensive review of online data repositories that could house data from the NEPs. Rather, the focus was on several federally supported domain-specific ocean and coastal acidification and water quality repositories that are recognized for meeting the requirements for FAIR[1] (Findable, Accessible, Interoperable, Reusable) data principles.

In this report, general best practices are described for choosing a data repository, using discovery tools, preparing metadata and data, understanding processes for submitting data, and accessing a published dataset. Overall, new collaboration and online data sharing tools and approaches offer greater opportunities for scientists and managers to communicate results to a broad community, including large continuous monitoring datasets. Because this report may introduce new terminology to the reader, a glossary is provided in Appendix A.

[1] *Fair Data Principles - https://www.go-fair.org/fair-principles/*

Project Overview

Estuaries and coastal areas are highly vulnerable to the impacts of ocean and coastal acidification (OCA), particularly on shellfish, coral reefs, fisheries, and the commercial and recreational industries that they support (Washington State Blue Ribbon Panel on Ocean Acidification, 2012; Gledhill et al., 2015; Saba et al., 2019). To assess this vulnerability, high resolution monitoring data are needed at varying spatial and temporal scales to provide actionable information tailored to each estuary. Estuarine specific drivers can contribute to acidification, such as nutrient enrichment from stormwater, agriculture and wastewater discharges, upwelling of CO_2-rich seawater, elevated atmospheric CO_2 from urban and agricultural activities, benthic and marsh-driven processes, and alkalinity and carbon content of freshwater flows (Duarte et al., 2013; Rheuban et al., 2019; Turner et al., 2021). Therefore, synoptic water quality surveys are expanding in scope to encompass carbonate chemistry parameters, such as pH and partial pressure of carbon dioxide (pCO_2). Since 2015, the Environmental Protection Agency (EPA) has funded and supported efforts to expand coastal acidification monitoring at several National Estuary Program (NEP) sites. The NEP is an EPA place-based program to protect and restore the water quality and ecological integrity of “estuaries of national significance.” Each of the 28 NEPs sets priorities and develops a plan for restoring and protecting water quality in their estuary and watershed. With partner organizations, NEPs conduct water quality monitoring to track trends and to evaluate changes in response to management actions. The NEP sites conducting coastal acidification monitoring using autonomous pH and pCO_2 sensors include Casco Bay (Maine), Massachusetts Bay (Massachusetts), San Francisco Bay (California), Santa Monica Bay (California), Tampa Bay (Florida), Long Island Sound (Connecticut/ New York), Coastal Bend Bays (Texas), Tillamook Bay (Oregon), and Barnegat Bay (New Jersey), Mobile Bay (Alabama), Indian River Lagoon (Florida), and Puget Sound (Washington).

These NEPs and their partners have generated a large volume of high temporal resolution carbonate chemistry data and associated parameters including temperature, salinity, and dissolved oxygen. These data have improved our understanding of spatial and temporal variability of coastal water carbonate chemistry and the drivers responsible for this variability (Rosenau et al., 2021). Several of the NEPs have telemetry capabilities for transmitting sensor data remotely in real-time. Other NEPs manually download the sensor data from an onboard instrument data logger during routine maintenance, typically every two to six weeks. Some sensors take measurements at 15-min intervals, which results in data files that can be larger than traditional water quality sampling programs. In some cases, observations are averaged and reported on an hourly basis to make the data easier to manage, but may result in the loss of detail and resolution. Detailed information about coastal acidification monitoring deployments in the NEPs can be found in EPA’s report: Measuring Coastal Acidification Using *In Situ* Sensors in the National Estuary Program (EPA, 2021).

The NEP OCA sensor monitoring projects have succeeded in collecting several years of high-resolution baseline data and are in the process of reporting results to the scientific and coastal management communities. NEPs, however, are frequently challenged by the capacity to store and provide access to high resolution continuous monitoring data. They typically do not have a consistent approach for analyzing, sharing, and archiving their ocean and coastal acidification data. Most of the data are stored on local computer networks at the NEP or with their partners, which include universities, state agencies, and other organizations. This has resulted in some data that are not as easily accessible (i.e., “findable”) to the broader scientific and management community. It is noted that some NEPs have developed partnerships with federal agencies such as U.S. Geological Survey (USGS) and NOAA’s Integrated Ocean Observing

Systems (IOOS) to make their data more accessible.

In addition to the NEPs, there are an increasing number of organizations, including community science organizations, that are beginning to monitor OCA data. The recently released EPA document "Guidelines for Measuring Changes in Seawater pH and Associated Carbonate Chemistry in Coastal Environments of the Eastern U.S" (Pimenta and Grear, 2018) provides guidelines for sample collection, preservation, and analysis. Many of these organizations, however, are also challenged by data management. As efforts and interests on OCA continue to rise around the nation, monitoring groups need information about how to manage data locally and make it more accessible to a wider public. In addition, there are new requirements from funders, agencies, and journal publishers to improve the accessibility of data so that it is available to the scientific community and other stakeholders. Some publications, such as *Nature*, recommend that where possible datasets should be submitted to discipline-specific, community-recognized repositories.[2] Therefore, EPA commissioned this report to provide information that NEPs, their partners, and other monitoring groups can use to submit time-series data to publicly accessible online data repositories.

Data repositories provide the capacity to store and provide public access to data which can minimize the in-house technology burden on individual organizations. By putting data into a data repository, the data and associated metadata collected can be made available for discovery and use in research, management, and industry operations. Storage of data in a repository can also satisfy data management and data sharing requirements from funding agencies and peer-reviewed publications. It is also important to note that sharing data in online repositories is a component of the trend towards Open Science[3] principles, where science is conducted in a way so that others can collaborate and share research data and processes (e.g., code for producing graphics or statistical analyses on platforms such as GitHub). However, finding the right repository for a dataset can be a time-consuming process.

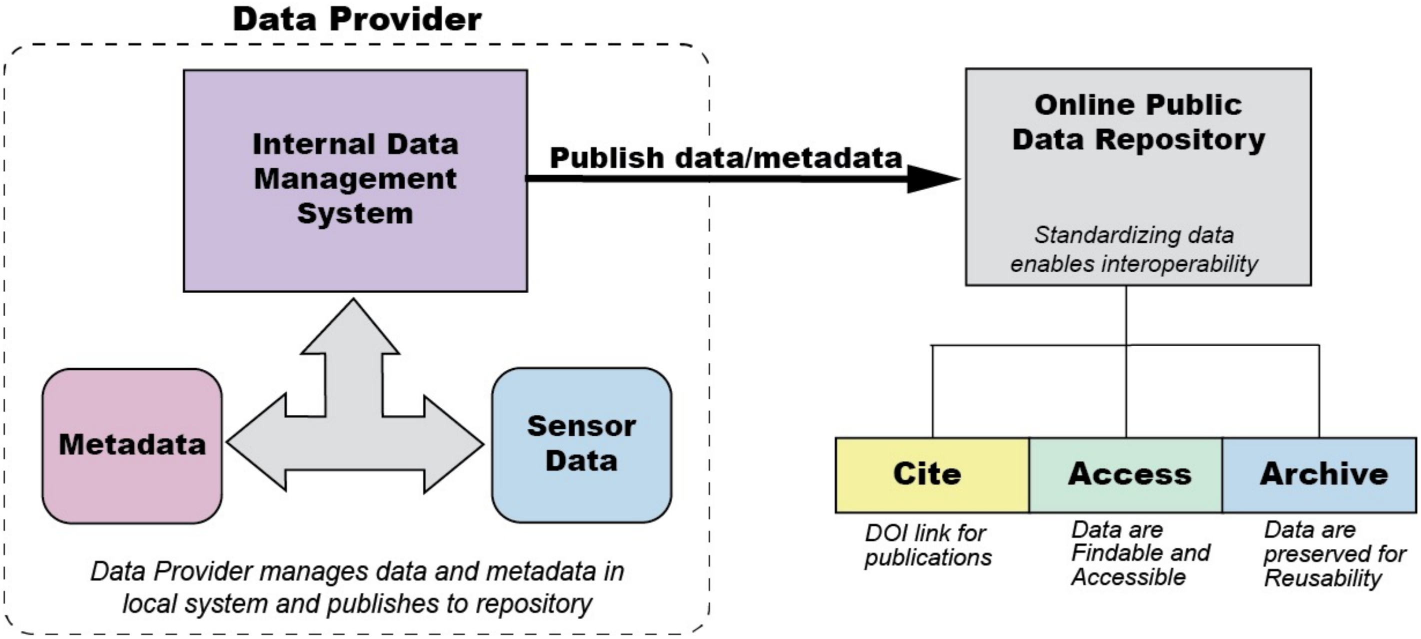

Data Provider
Internal Data Management System
Publish data/metadata
Online Public Data Repository
Standardizing data enables interoperability
Metadata
Sensor Data
Data Provider manages data and metadata in local system and publishes to repository
Cite
Access
Archive
DOI link for publications
Data are Findable and Accessible
Data are preserved for Reusability

Figure 1 - Adapted from Common data management workflow (Amorim, 2014). This figure demonstrates a common data workflow where the data provider (e.g., NEP project data manager or principal investigator) collects sensor data, prepares metadata, and stores both records together in an internal data management system. By publishing data to an online repository, these datasets become findable and accessible to a larger audience of data users.

[2] https://www.nature.com/sdata/policies/repositories

[3] Center for Open Science: https://www.cos.io/

Fully reviewing and evaluating the many publicly available data repositories would be a large undertaking outside the scope of this report. At the outset of this project, a few key repositories were identified that were deemed by the authors and advisors to be well-suited for the continuous coastal acidification monitoring data collected by the NEPs. Although this report focuses on the needs of the NEP and their partners, additional audiences that could benefit from this project include academia, the Coastal Acidification Networks, non-governmental organizations, and citizen and community science groups involved in collecting OCA data (e.g. Gassett et al., 2021). In addition, the best practices described are applicable to other water quality monitoring programs.

Background on Online Data Repositories

Online data repositories are large, well-managed, database infrastructures designed to store data, enable public access to datasets, and in some cases act as a permanent archive for the data. Repositories are a central common point for storing related files and may be specialized by subject matter, scientific discipline, or geospatial region. All repositories support the idea that sharing data improves awareness of the results of a study and enables use of data beyond the initial purpose. Making research and monitoring data available online increases transparency, supports reproducibility of the original work, and enables data users anywhere to access, share, understand, compare, and synthesize results from the research or monitoring effort.

Data repositories vary in the level of complexity and required elements from the data provider. At one end of the spectrum are the generalist repositories that are open to submissions of data regardless of data type, content, or discipline. These repositories have minimal requirements for metadata that are necessary to catalog the dataset for discovery through search tools and allow the dataset to be cited by others (i.e., citation metadata). At the other end of

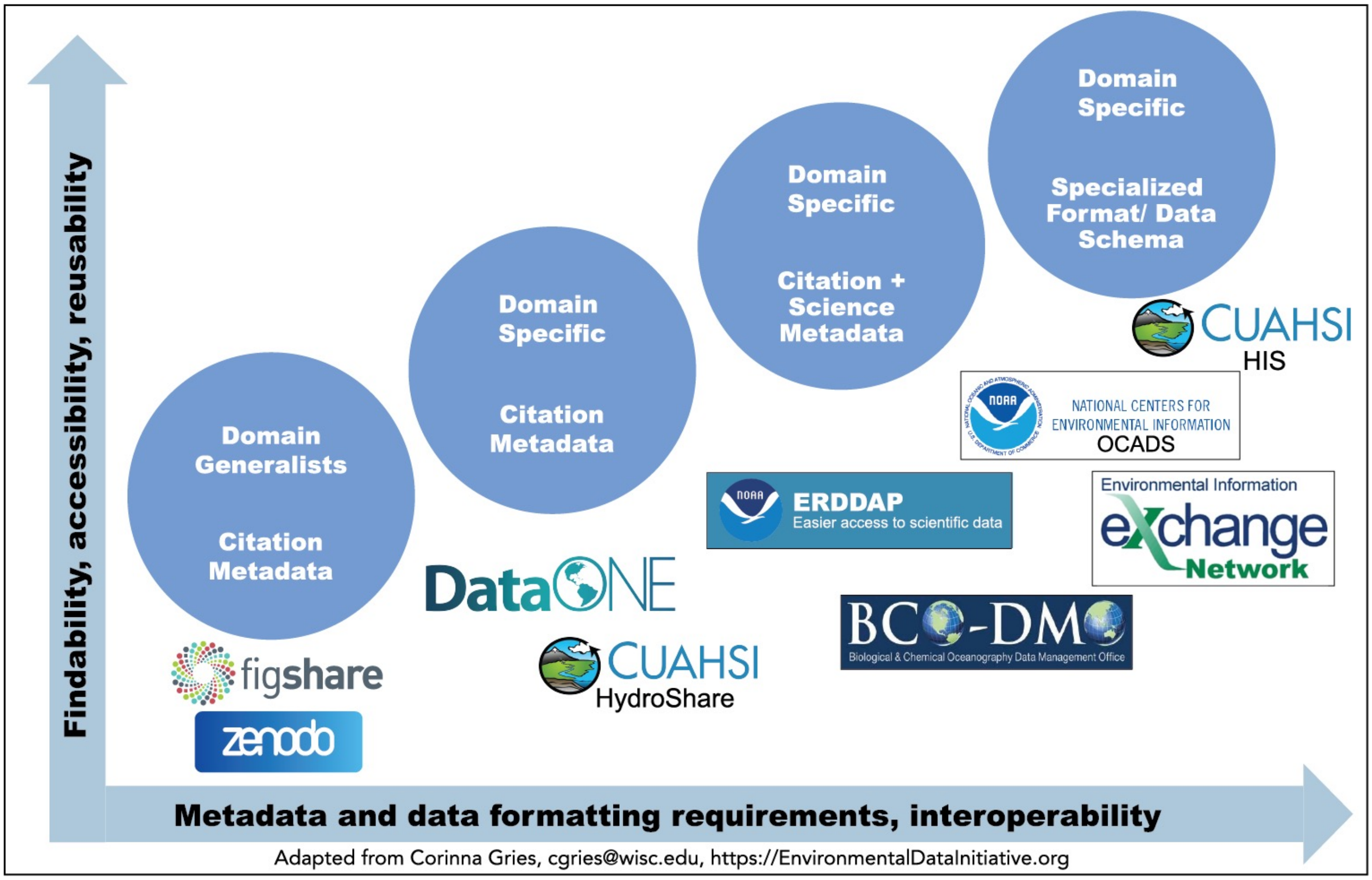

Findability, accessibility, reusability
Domain Generalists
Citation Metadata
Domain Specific
Citation Metadata
Domain Specific
Citation + Science Metadata
Domain Specific
Specialized Format/ Data Schema
figshare
zenodo
DataONE
CUAHSI HydroShare
NOAA ERDDAP Easier access to scientific data
BCO-DMO Biological & Chemical Oceanography Data Management Office
NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION OCADS
Environmental Information eXchange Network
CUAHSI HIS
Metadata and data formatting requirements, interoperability
Adapted from Corinna Gries, cgries@wisc.edu, https://EnvironmentalDataInitiative.org

Figure 2 - Adapted from figure developed by Corinna Gries, Environmental Data Initiative https://EnvironmentalDataInitiative.org. This figure illustrates the range of data repositories from the general to the highly specialized. Repositories in the bottom left have fewer requirements for the data provider while the repositories in the upper right have specific requirements and data formats.

the spectrum are repositories that are discipline-specific and require metadata and datasets to conform to a specific data schema that defines variable names and units, often with a specific, standard vocabulary. The repositories evaluated in this project were those that are domain-specific with more general metadata requirements.

While some repositories may fit somewhere in between the two ends of the spectrum, a citation metadata data repository allows the data provider to upload data in the original (or native) file format (e.g., .csv, Excel, multidimensional space-time data sets (NetCDF)) along with metadata required by the individual repository. These metadata typically include basic descriptive information (e.g., study name, data provider name, institution, temporal and geospatial bounds of study, project abstract, keywords) and information about the fields (e.g., parameters, units, calculations). Once datasets are added to the repository, they can be found using search tools that query the metadata and the original data files can be downloaded by end users for further analysis. Some repositories enable users to organize similar datasets into collections around specific studies or events.

The data repositories that have a specialized data schema are more complex but enable greater interoperability of datasets. These data repositories require that datasets (generally time series) conform to a defined data schema with required fields vocabularies for naming conventions the data itself (e.g., parameter name, units) that must be matched by the incoming dataset. Templates are available to the data provider to map or align the metadata and data fields to the data schema. From there, the datasets can be integrated into a relational database where the data observations can be queried, visualized, and compared to other data using tools provided by the repository.

In all types of repositories, well-described datasets are more findable and accessible to the public. A citation metadata data repository can be an easier path for a group with limited time or resources for data management because of the minimal metadata requirements and the ability to submit data in the original file format. Additionally, in a citation metadata data repository, the data provider can define the parameter names and units for the data. By contrast, the data repositories with a more defined data schema and controlled vocabularies require that the data provider transform the data to conform to the data schema.

Approach

To better understand the OCA data management needs of the NEPs, a series of interviews were conducted with NEP data providers and users of the data. Data providers often include the principal investigators, or technical data managers for the project, who are responsible for maintaining datasets on local computer networks. Data providers often report the analyses of data to the public, such as a NEP State of the Bay report. Interviews with data users outside of the individual NEPs were also conducted to help inform NEP principal investigators and data managers about additional uses of OCA data and considerations about documenting and sharing these data.

The findings from the interviews were synthesized to produce a set of attributes to evaluate the data repositories. These attributes were then used to inform three test cases that were conducted with data providers (who are also principal investigators of the research) to evaluate the process of preparing and submitting data to online repositories. This report summarizes the development and application of the attributes in the test cases including: 1) identifying and selecting a data repository; 2) preparing metadata and data for submission; and 3) evaluating accessibility of the data. Online data repositories are identified that are suitable for submission of continuous ocean and coastal acidification data collected by the NEP and other monitoring groups.

Lastly, general best practices are described for accessing and using search tools in a data repository, preparing metadata and data, understanding processes for submitting data, and testing for access and publication of a dataset.

Evaluation of Online Data Repositories

Interviews with Data Providers and Data Users

After reviewing EPA's report Measuring Coastal Acidification Using *In Situ* Sensors in the National Estuary Program (EPA, 2021), five data providers were interviewed to better understand how NEPs are managing and sharing OCA data and submitting data to online data repositories. Three data users were also interviewed to identify issues related to accessing data from repositories or issues working with the OCA data in general.

Data Provider Interviews: Submitting Data to a Repository

Interviews were conducted with five data providers regarding the workflow from data collection to data submission. They were asked about the type of data being collected; the workflow from the sensor to the internal data management system; if metadata records had been established; quality assurance/quality control (QA/QC) processes; whether the NEP had submitted or planned to submit to a public data repository; attributes for selecting a repository; and challenges or concerns submitting data to that repository. Full questions and responses are provided in Appendices B and C.

Common practices and issues emerged. The NEPs and their monitoring partners:

- manage sensor OCA data in-house (database, local file store) and make data available upon request (but not online);
- establish metadata records that are not always complete;
- can share data with stakeholders including state agencies, scientists, and more recently aquaculturists;
- have technical capability to manage data and share to a public repository, but lack time to evaluate and choose repository and maintain regular updates (e.g., quarterly, annually);
- are interested in making OCA data more accessible;
- are interested in more guidance on QA/QC for OCA data; and
- are interested in common processes and tools, such as the NOAA National Estuary Research Reserve System's Centralized Data Management Office (CDMO).

Data User Interviews: Accessing Data in a Repository

Interviews were conducted with three users of OCA data about their experiences discovering and accessing data from online repositories. These users included a university research scientist interested in OCA data to support research and modeling efforts, and two NEP managers who use OCA data to produce synthesis reports on local environmental conditions (e.g., Casco Bay State of the Bay report, 6th edition). The full list of questions and responses are provided in Appendices B and C. Common issues are grouped into three categories – data discovery, metadata, and data formats.

Data Discovery

Data users often must look at multiple repositories, which all have their own data access and discovery tools. Common repositories mentioned were NOAA's National Centers for Environmental Information (NCEI) Ocean Carbon and Acidification Data Portal (OCADS), Biological & Chemical Oceanography Data Management Office (BCO-DMO), and state agency data repositories. The larger online repositories generally have good search tools and return datasets that match the search queries. However, the results often return too many datasets in the search results, requiring the user to evaluate each result. The user must often repeat the search with different keywords and parameters to further filter the list or evaluate each data set individually. The repositories generally organize the datasets by title which can be complicated and hard to understand. Often, the data user must contact the data provider to answer specific questions before they feel comfortable using the dataset. An ideal functionality for a repository, as identified by the data users interviewed, would include better search functionality

and the ability to programmatically acquire the data through web services such as an application programming interface (API).

Metadata

While the data users had different purposes for the data, the most common issues were that the metadata provide insufficient information for a dataset. There is often inconsistent use of field names by data providers measuring similar parameters as well as a lack of detail about accuracy and uncertainty ranges of sensors. Data providers have different styles and approaches for describing metadata, and data users find that they often need to search through protocols to find the relevant information. Some data repositories generate search lists from user-entered parameters which can result in multiple versions of similar terms making it difficult to do effective searches.

The QA/QC processes are often not well described in the metadata or are buried in protocols. Data users found it difficult to understand processes around flagged data. More clarity on how data were flagged, definitions of the flags, and whether the flagged data were removed from the dataset would be desired. To avoid making assumptions, data users regularly need to contact the data provider to address questions about the data. Data users found that they are more likely to use data when the metadata are clear and QA/QC processes are well described.

A basic metadata record that would satisfy the data user needs would include:

- well-documented QA/QC procedures, ideally represented in formal documentation of data quality practices such as a Quality Assurance Project Plan (QAPP);
- information about methodology and study design that is easy to find;
- description of how data were processed or analyzed;
- well-described variables using common vocabularies for terms (e.g., units, instruments, date/time, location); and
- detail on how flagged data are handled.

Data Formats

There were differences in preferred file formats among the data users. Some prefer NetCDF output while others prefer .csv/Excel if well-described. Some common problems that were noted with .csv and Excel files had to do with file size and file structure. When the data files are too large, it can create issues with analysis programs (e.g., R). In cases where data providers add metadata to the header of the document, it can cause problems with analysis programs skipping rows.

Descriptions of Online Data Repositories

Data Repositories Considered

Working with project partners at the outset of this project, several large, federally supported repositories with a focus on water quality data and/or OCA data were selected for evaluation.[4] These include:

- NOAA's National Centers for Environmental Information (NCEI) Ocean Carbon and Acidification Data Portal (OCADS)
- The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI)'s HydroShare Data Portal and Hydrologic Information System (HIS)
- EPA's The Exchange Network (EN)
- NOAA's Integrated Ocean Observing System (IOOS)
    - Specifically, IOOS Regional Associations that are using ERDDAP[5] software to manage and disseminate data.

These data repositories were reviewed and evaluated based on the issues identified during the interviews with data providers and data users. The evaluations outline the challenges and opportunities associated with submitting data to different repositories. The repositories were also evaluated on the degree to which they adhere to the FAIR data standards, which include making data Findable, Accessible, Interoperable, and Reusable (Wilkinson et al, 2016). While this evaluation approach was specifically applied to a few select repositories based on the type of data, the structured approach described here could be used to evaluate other data repositories.These repositories are briefly described below. Additional information about preparing data and submitting to the repositories is available in **Appendix D**.

National Centers for Environmental Information
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Name: NCEI Ocean Carbon Data System (OCADS)

URL: https://www.ncei.noaa.gov/access/ocean-carbon-data-system/

Owner/operator: NOAA National Centers for Environmental Information (NCEI); with funding provided by NOAA/Climate Program Office/Ocean Observing and Monitoring Division, NOAA/Ocean Acidification Program, and the National Aeronautics and Space Administration (NASA).

NOAA's NCEI OCADS is a data repository established in 2017 as a carbon data specific repository with a mission to host and provide access to ocean carbon datasets collected worldwide. It is also the permanent home of data from the Carbon Dioxide Information Analysis Center (CDIAC), which includes over 30 years of data, as well as newly acquired ocean carbon data. The data are stored and served from NCEI's archive data access services. NCEI supports FAIR data publication and is listed in the Repository Finder tool https://repositoryfinder.datacite.org/ developed by the Enabling FAIR Data Project. The OCADS data repository specializes in carbon data and includes templates with metadata and data recommendations for parameter names and attributes specific to OCA. The template was developed by working with the OCA research community to their meet scientific needs. OCADS manages controlled vocabularies for many metadata elements such as parameters, observation type, instruments (including uncertainty descriptions specific to certain sensor manufacturers), institution type, etc. All parameter names provided in the dataset are mapped to a controlled vocabulary; previously this was done by NCEI after submission. Going forward, data providers will be able to do this during submission using the Scientific Data Information System (SDIS).

[4] *Other National Science Foundation (NSF) supported repositories such as Biological & Chemical Oceanography Data Management Office (BCO-DMO) for ecosystem research data and DataONE were also considered but this project was limited in scope.*

[5] *https://coastwatch.pfeg.noaa.gov/erddap/index.html*

Name: CUAHSI HydroShare

URL: https://www.hydroshare.org/

Owner/operator: The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI)

HydroShare is a data repository developed by CUAHSI for the purpose of storing water-resources data and models, and making them accessible to the public. HydroShare provides a data portal to discover data as well as tools to manipulate and visualize data. HydroShare was developed and supported by the National Science Foundation (NSF) as well as an active membership program. It was built using open-source software (Python/Django, iRODS) and has a collaborative project on GitHub (https://github.com/hydroshare/) to further develop the platform and share data and models. CUAHSI supports FAIR data publication and is listed in the Repository Finder tool https://repositoryfinder.datacite.org/ developed by the Enabling FAIR Data Project.

The data types supported include hydrological time series, geographic features (vector data), geographic rasters (gridded data), multidimensional space-time data sets (NetCDF) and composite/complex datasets (river geometry).

HydroShare only requires basic citation metadata for submitted datasets. The actual datasets submitted to HydroShare are not automatically integrated into a relational database where the values can be queried or visualized through tools on the website. The required descriptive metadata that accompanies the dataset enables discovery through the search portal. HydroShare enables collaboration during the preparation of a dataset and provides capacity to group individual datasets under a common theme or project. Metadata entries can also be created on HydroShare where the dataset itself are not uploaded, but a link is provided to access data stored in another location.

Name: CUAHSI Hydrologic Information System (HIS)

URL: http://data.cuahsi.org/

Owner/operator: The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI)

CUAHSI's HIS portal, known as HydroClient, provides access to millions of time series observations collected by federal agencies, researchers, and volunteer groups. The data are primarily hydrological and include stream gauge measurements, meteorological stations, grab samples, and soil moisture measurements. It is a more complex data repository that requires that metadata and data are transformed into a discipline-specific data schema for hydrological data where they are then integrated into a relational database.

Data submitted to HIS are available through the HydroClient data portal. Datasets can be directly downloaded in the original format or visualized through online tools available on HydroClient.

Upon discussion with CUAHSI program staff, the newer HydroShare data tool was recommended for time series such as the continuous OCA sensor data. This was largely because the template required to map datasets for submission to HydroClient can be cumbersome and the recommended format is specific to the hydrological community.

Name: The Exchange Network (EN)

URL: https://www.epa.gov/exchangenetwork, https://www.exchangenetwork.net/

Owner/operator: U.S. Environmental Protection Agency

EN was designed to improve data sharing between states, territories, tribes, and EPA. Prior to the EN, data sharing was difficult because of inconsistencies with how data were collected and stored as well as incompatibility with computer systems. The EN consists of community developed "exchanges" that are essentially data schema for specific disciplines. Generally state agencies adopt relevant schema data standards into their own data systems and set up EN software (known as a server node) that enables them to share data directly to EPA's Central Data Exchange through standardized web services and APIs. For water data, the Water Quality Exchange (WQX) is the primary data exchange. Non-state agency data providers can share data directly with the relevant state agency through individual arrangement. Non-state entities can use a tool called WQX Web (https://www.epa.gov/waterdata/wqx-web-account-registration) to upload data files. The process does involve setting up a user account for the organization and mapping the data to the WQX format. WQX currently does not support continuous water quality monitoring data and therefore it may not be the ideal repository for the NEP OCA data. However, EPA continues to explore approaches for sharing and storing continuous monitoring data. If high frequency time series such as continuous monitoring data is submitted to the Exchange Network, it is recommended that data be summarized by hourly or daily averages with the full data logger files uploaded as an attachment.

Name: The U.S. Integrated Ocean Observing System (IOOS) Regional Associations

URL: https://ioos.noaa.gov/regions/

Owner/operator: NOAA

IOOS is a coordinated network of people and technology that work together to compile and distribute observation data from ocean and coastal environments and guide the development of activities within defined regions. IOOS is comprised of eleven Regional Associations (RAs) that serve stakeholders from the nation's coastal communities including the Great Lakes, the Caribbean and the Pacific Islands and territories. NEPs can partner with IOOS RAs to make OCA data accessible through the RA's data distribution services. Most of the eleven IOOS RAs have implemented a data management software tool called ERDDAP to provide access to a multitude of ocean observing datasets from the region. For example, the San Francisco Estuary Partnership works with the Central and Northern California Ocean Observing System (CENCOOS) to integrate data from their OCA monitoring buoy. The data are available through the CENCOOS ERDDAP installation https://erddap.cencoos.org/erddap/tabledap/tiburon-co2.html.

All RAs have a charge from IOOS to serve as a Regional Information Coordination Entity (RICE)[6]. This includes integrating local ocean observing data from the region and making it accessible to the public. The first step for an interested NEP partner is to reach out to the regional RA to discuss the data and develop a plan for integrating the data.

[6] https://ioos.noaa.gov/about/governance-and-management/certification-extending-reach-regional-data/

Online Data Repository Evaluation

A series of attributes were developed from the results of the interviews to evaluate the individual data repositories. The attributes were organized into four categories: metadata, data preparation and submission, data access, and other attributes. Other attributes include information about long-term stability of the repository, associated costs, data archiving, and additional notable features. The following questions were developed for each category:

Metadata

- Are there metadata requirements?
- Are there templates provided?
- What metadata standards are used?
- Are there metadata file format requirements?
- Do the repositories adhere to FAIR principles?

Data Preparation and Submission

- What guidance is available for data providers?
- Are data templates available?
- What are the file type requirements/recommendations?
- What is the process for submitting data?

Data Access

- How are data accessed?
- What guidance is provided for users?
- What filter/tools are available?
- Are there other OCA datasets in the repository?
- Are there APIs for data?

Other Attributes

- Are Digital Object Identifiers (DOIs) available?
- Is the repository expected to be stable in the long-term?
- Are there collaborative features?
- Is there a cost to submit and publish data?
- Does it meet the FAIR data standard?

Metadata

All the repositories required some level of metadata that describe the dataset and are used to find the dataset through the search tools. OCADS required additional metadata that describe the carbon data parameters being collected. CUAHSI HIS and EPA's EN required that all metadata conform to a discipline specific schema. All repositories provided support for preparing metadata for submission through examples or templates. There were varying levels of guidance or requirements to use specific discipline-based vocabularies or standards for the data. While the initial repository evaluation occurred before the test case work, it should be noted that this was an area that was identified as needing the most time to navigate by the test case participants.

TABLE 1. METADATA EVALUATION ATTRIBUTES

	DATA REPOSITORY				
	OCADS/NCEI	CUAHSI HydroShare	CUAHSI HIS	IOOS ERDDAP	EPA Exchange Network
Metadata Requirements	Yes	Yes	Yes	Yes	Yes
Metadata Templates Available	Yes	Yes	Yes	Yes	Yes
Metadata Vocabularies	WHP-Exchange Format (CCHDO); WOCE flags recommended	No specific vocabularies or data standards required	Dublin core, ODM, data schema, and controlled vocabulary	CF	WQX (Water Quality data Exchange) schema and controlled vocabulary
Metadata Submission Process	Metadata added through form on SDIS site	Citation metadata through form, can upload supplementary PDF	Added to Excel templates (Standard or Advanced)	Work directly with Regional Association; ERDDAP has data submission template	Work directly with Exchange Network node (usually state agency) or submit via WQX Web

CF – NetCDF Climate and Forecast (CF) Conventions
CCHDO – Clivar and Hydrographic Data Office (for global CTD and hydrographic data)
ODM – Observations Data Model, data structure used by CUAHSI HIS
WHP Exchange format – a text-based format for bottle and CTD data
WOCE – World Ocean Circulation Experiment

Preparing/Submitting Metadata and Data

The evaluation attributes for preparing and submitting data included whether metadata and data templates were available, if there were data type limitations (e.g., numeric, date/time, spatial, string), recommendations or requirements for file format, and guidance on how to submit data.

TABLE 2. EVALUATION ATTRIBUTES FOR PREPARING AND SUBMITTING DATA

	DATA REPOSITORY				
	OCADS/NCEI	CUAHSI HydroShare	CUAHSI HIS	IOOS ERDDAP	EPA Exchange Network
Data Templates Available	Yes - additional data templates for underway, profile, mooring	Yes - also provides examples of other datasets	Yes - template for data schema	Yes - IOOS ERDDAP gold standard	Yes - template for data schema
Dataset File Format	ASCII/.csv, NetCDF	.csv, .xlsx, PDF	.xlsx	ASCII/.csv, NetCDF, XML	XML, .xlsx, .csv
How to Submit Data	Create account on OAP Science Data Information System (SDIS) website	Create account, use website form to provide metadata, and upload dataset	Fill out data templates, upload files through website	Work directly with Regional Association; ERDDAP has data submission template	Work directly with Exchange Network node (usually state agency) or submit via WQX Web

Accessing Data

Evaluation of attributes for accessing data included description of the data access portal, availability of an API for acquiring data, ease of discovery/access of data, and whether other OCA data are in the repository. All the repositories have data portals with similar features for discovering and accessing data. The queries can be done by filtering based on keywords, location, and date. The query results are only as good as the data entered by the data provider, reinforcing the need to develop and provide good, descriptive metadata.

TABLE 3. EVALUATION ATTRIBUTES FOR ACCESSING DATA

	DATA REPOSITORY				
	OCADS/NCEI	CUAHSI HydroShare	CUAHSI HIS	IOOS ERDDAP	EPA Exchange Network
Data Access Portal	Data Portal - map/filter search	Data Portal - map/filter search	Data Portal - map/filter search	ERDDAP search/ filter interface, API	USGS/EPA Water Quality Portal map/ filter search
API for Data	Yes - for metadata, data stored in files (.csv/.xlsx)	Yes - for metadata, data stored in files (.csv/.xlsx)	Yes	Yes - for metadata and data	Yes - for metadata and data; data can be combined across organizations because the data are in common format and use common terms
Ease of Accessibility	Filter based search (keywords, location, date)	Filter based search (keywords, location, date)	Filter based search (keywords, location, date)	Filter based search (keywords, location, date)	Filter based search (keywords, location, date)
Data Discovery URL	https://www.ncei.noaa.gov/access/oads/	https://www.hydroshare.org/search	https://data.cuahsi.org/	Varies by institution, CENCOOS example: https://erddap.cencoos.org/erddap/tabledap/tiburon-co2.html	https://www.waterqualitydata.us/
Other OCA Data in Repository	Yes - OCA data are the primary focus of repository	Yes	Yes	Yes - varies by RA	Yes

Other Attributes

The attributes for this category includes issues raised in the interviews focused on publishing and collaboration not covered by the previous categories. They included the ability to get a Digital Object Identifier (DOI) number, long term storage or archiving policies, collaborative features, costs, versioning, limitations, and stability of platform.

TABLE 4. OTHER EVALUATION ATTRIBUTES

	DATA REPOSITORY				
	OCADS/NCEI	CUAHSI HydroShare	CUAHSI HIS	IOOS ERDDAP	EPA Exchange Network
DOIs Available	Yes - Upon request, are available after data are accepted by NCEI (1-2 weeks)	Yes - DOI numbers are minted immediately on publishing data	Yes - DOI numbers are minted after data are accepted	Not issued by ERDDAP but can be added as variable in configuration	No
Long Term Data Storage	Yes - Data are archived at NCEI (NOAA's official data archive)	Yes - NSF commitment to maintain data repository	Yes - NSF commitment to maintain data repository	No - but RA may have an archiving plan with NCEI	Yes - Water Quality Portal
Collaborative Features	No	Yes - can group datasets under collections, tools for interaction within platform	No	No	No
Cost for Submitting	Not for NOAA supported data; May be costs for very large datasets; Non-NOAA funded data will be subject to approval	No cost at present and no immediate plans to change	No cost at present and no immediate plans to change	RA may request support for publishing data	No
Data Publishing Options	Submitted data are publicly available	Submitted data can be kept private until ready to make public	Submitted data are publicly available	Public only	Supports public and private data (e.g., regulatory)
Limitations	Datasets larger than 2 GB must be sent via FTP; for larger datasets, contact OCADS directly	2 GB data per user but can create additional users for collection if needed. Larger datasets can be added directly to the party middleware used by CUAHSI (iRODS)	2 GB data per user. Larger datasets can be added directly to the party middleware used by CUAHSI (iRODS)	Work with Regional Association to set up agreement before submitting	WQX is not structured for continuous monitoring data, requires conversion to binned data
Long Term Stability	Yes - Uses NCEI data system as backend. OCADS is upgrading to better metadata user interface and search portal in 2021	Yes - Data system stored at Renaissance Computing Institute (RENCI), supported by NSF; offsite backup storage	Yes - Data system stored at Renaissance Computing Institute (RENCI); supported by NSF; offsite backup storage	Yes - RAs are supported by NOAA IOOS; specific data systems can vary by RA	EPA has long-term commitment to EN

Data Repository Test Cases

Following the initial review and evaluation of the data repositories, test cases were conducted to determine which repository may be the best fit for the data management needs of the NEPs and related OCA monitoring groups.

Test Case: Shell Day Data
https://iopscience.iop.org/article/10.1088/1748-9326/abcb39; http://www.necan.org/ShellDay

Data Provider: Jennie Rheuban, Woods Hole Oceanographic Institution

Data Summary: Citizen Science data from Shell Day monitoring event. Discrete samples of surface water samples were collected up to five times at different tidal stages on one day throughout New England in August 2019. Bottles were delivered to participating laboratories and analyzed for total alkalinity. Salinity and temperature were also recorded by samplers.

Data Description: Discrete sampling data/historic/final data set

Requirement for Data Repository: Need to submit data to accompany a paper in publication, DOI is required.

Repository Tested: CUAHSI HydroShare

Recommend: Yes

Link to Dataset: https://www.hydroshare.org/resource/4364cffedc7e49d49255eef5f8e83148/

Overall

Ms. Rheuben reported that the HydroShare platform was straightforward and easy to use, largely because she had a separate file for a detailed metadata record and the data files were ready to submit. It only took her a few hours to set up an account, fill out the required metadata fields through an online form, and upload the individual data and supplemental metadata files. While developing the record, she noted that the data can be kept private, and other collaborators can be given access. This was extremely useful in finalizing the dataset with her collaborators.

Metadata

General project metadata elements were required and included project name, authors, keywords, project summary, and temporal and spatial boundaries. Additional fields were available to list contributors and funders. There is no requirement or recommendation for structuring the data in terms of vocabulary or required parameters. The general project metadata elements included an abstract, or description, of the dataset. She felt this was easy to generate using content from the paper that was written about the project and dataset. Additional sentences were added to describe the parameters and keywords.

Ms. Rheuben had familiarity submitting other carbonate datasets to OCADS where additional metadata requirements include information about the data. These metadata are mapped to a template that allows for information to be provided about uncertainty for carbonate data parameters. She was surprised that HydroShare did not provide a similar template.

Submitting Data

There was flexibility on the platform to use a variety of data formats for data and metadata files which was helpful. There are several states of data visibility after submitting data that are well-described. These include Private - not visible publicly, but can be shared with collaborators; Public - publicly visible, but DOI isn't published, can still edit; and Published - DOI finalized, dataset record is closed to edit.

Publication and Collaboration Tools

The feature to add collaborators for co-editing in draft mode was helpful to Ms. Rheuben and is a feature unique to HydroShare. For this highly collaborative project, she found it useful to have the ability to add co-authors and funders. However, the format was structured for federal grants and didn't fit all funder cases such as foundations.

Before adding the keywords, she tried to infer what the search tools are looking for to tailor the information so the dataset would be discoverable. She looked at a few other datasets and did something similar with the structure of her metadata.

Getting a DOI quickly was a requirement for this data set as a publication was about to go to press. The DOI was minted instantly, and a citation was provided in draft mode even though the data weren't finalized and published. This was particularly helpful for her to be able to send to the journal immediately.

There were some lingering questions Ms. Rheuban had about best practices for updating this dataset annually. It wasn't clear to her if it is recommended to do a separate submission with a new DOI issued each year. But she noted that approach is consistent with projects like the Global Carbon Project that has a yearly citation and DOI for data.

Casco Bay Estuary
PARTNERSHIP

Test Case: Casco Bay Estuary Partnership Continuous Monitoring OCA system, South Portland, Maine

Data Provider: Dr. Chris Hunt, University of New Hampshire (UNH)

Data Summary: Casco Bay Estuary Partnership is an NEP collecting continuous monitoring OCA data. The University of New Hampshire manages the sensors and processes the data.

Data Description: One continuous monitoring station, housing sensors for measuring pH, salinity, temperature, pCO_2, and DO, was deployed in Casco Bay. There is no telemetry, so once per month the data were downloaded from the sensor loggers. The raw data were processed by UNH using MATLAB and vendor-provided software and output to an Excel file with averaged hourly readings. The data were then organized in annual files and sent to the Casco Bay Estuary Partnership.

Requirement for Data Repository: None, however, it is planned to publish the data findings in the future.

Repository Tested: OCADS

Recommend: Yes

Link to Dataset: https://www.ncei.noaa.gov/data/oceans/ncei/ocads/metadata/0229832.html

Overall

Because of his familiarity submitting other datasets to NCEI, Dr. Hunt found the process for submission to OCADS was straightforward and easy to follow. The website provided clear guidance on accessing the metadata templates and how to submit once they are completed.

Metadata

Dr. Hunt stores the monthly data logger data in an Excel file that contains basic metadata about the dataset. Each year an annual report is submitted to the Casco Bay Estuary Partnership that contains additional metadata.

OCADS provides a metadata submission form (as an Excel file template) specific to OCA data with required and recommended fields. In the OCADS template, there are over two hundred metadata elements where information can be provided if available, but only about 36 fields are required.

The most specificity that is asked for in this template is for calibration information; for this dataset, that only applied to manufacturer information and was easy to add. For data that might have a lab analysis component or was measured on a research or monitoring cruise, there would likely be more information about calibration that could be added, and the template would accommodate that data.

Submitting Data

Once the metadata submission form was filled out, it was an easy process to email the contact person listed on the OCADS website. Dr. Hunt received confirmation that the metadata and data files were received, and he was given an estimate of about a week to have the data published. He received notice within a week that the data were published. Note: since the use case testing, NCEI has updated to the Scientific Data Information System (SDIS) for submitting data. More information is available in **Appendix D**.

Publication and Collaboration Tools

OCADS did not have the collaboration tools that were present in CUAHSI. Once submitted and accepted by OCADS, the data set cannot be edited, and all data are publicly available.

Friends of Casco Bay
Casco BAYKEEPER

Test Case: Friends of Casco Bay

Data Provider: Mike Doan, Friends of Casco Bay

Data Summary: Casco Bay, Maine

Data Description: A continuous monitoring station has been deployed off Cousins Island in Yarmouth, near the coastal midpoint of Casco Bay since 2016. Year-round, the station collects hourly measurements of depth, temperature, salinity, dissolved oxygen (DO), chlorophyll, turbidity, pH, and the partial pressure of carbon dioxide (pCO_2). The data are processed annually and shared with the Casco Bay Estuary Partnership.

https://www.cascobayestuary.org/resources/data-open-science/

https://www.cascobayestuary.org/casco-bay/monitoring/

https://www.cascobay.org/our-work/science/continuous-monitoring-station/

Requirement for Data Repository: None, however, there is interest in sharing data more broadly.

Repository Tested: CUAHSI HydroShare

Recommend: Yes

Link to Dataset: http://www.hydroshare.org/resource/0aadc8e61e68436abb5e99e0be6565a2 (*currently in private mode)

Overall

Friends of Casco Bay downloads the data from the sensor data logger every two to three weeks and manages the data in a local database along with metadata about the project and sensor. They also have a QAPP for the monitoring program that is available upon request. They have been interested in sharing data more broadly but have not submitted to an online data repository. The process to create an account on HydroShare was simple and straightforward.

Metadata

The basic metadata that is required to set up the dataset is added through a form on the website. Having this information ready to go made the process very easy. Looking at a few similar data records was helpful in structuring keywords and additional metadata for the data set.

Submitting Data

The data are kept in an Excel file and the metadata in a PDF. It was straightforward to upload the data as attachments to the record.

Publication and Collaboration Tools

The dataset is on HydroShare but is still in unpublished mode pending final decision from Friends of Casco Bay to make it publicly available.

Results

As a result of the interviews, online data repository review, and the test cases, two primary data repositories, NOAA's OCADS and CUASHI's HydroShare, were found to be suitable options for the NEPs and other data managers for submission of ocean and coastal acidification data (both discrete and continuous). These repositories both employ a similar approach to basic metadata requirements to make datasets more discoverable and accessible. HydroShare requires basic metadata that describes the project (e.g., who conducted the research, the temporal and geospatial bounds of the study, a summary or abstract description of the project, and keywords that are used to find the dataset). OCADS also requires similar project level metadata as well as metadata specific to OCA parameters (e.g., parameter name, units, uncertainty of sensor). The other repositories that were evaluated are also suitable for OCA data but had additional process steps that were not evaluated with the test cases. These repositories were CUAHSI HIS, EPA Exchange Network, and IOOS Regional Association partnerships.

The HydroShare platform is very easy to use and provides some useful features that OCADS does not. Specifically, HydroShare allows users to: collaborate on dataset submission, organize datasets under a collection, and select from several options for publication states based on the requirements of the data provider. One of the limitations of HydroShare, however, is a lack of specific requirements or prompts for detailed metadata about OCA parameters which puts the onus on the data provider to include that information.

NOAA's OCADS is a discipline-specific repository for global carbon chemistry data and is a valuable resource for the OCA community, providing a "one-stop shop" for OCA data. OCADS provides more structure to the data provider by requiring a metadata template to describe OCA metadata. While there are more required metadata fields than the other platforms, it is up to the data provider to create as thorough a record as possible. Submitting the data is straightforward, and the OCADS contacts were quick to respond once the data were submitted.

In sum, if the data provider values sharing additional datasets that are not specifically measuring OCA data, HydroShare would be suitable for the collaborative features and ability to group multiple datasets into collections. In contrast, if the data provider is only interested in submitting OCA data, the OCADS repository could be the best choice. OCADS is interested in being a clearinghouse for OCA datasets and synthesized data products such as the recently developed Coastal Ocean Data Analysis Product for North America (CODAP-NA) dataset comprised of data from over 60 cruises and 3000 oceanographic profiles[7].

There is also a hybrid model where an NEP could use HydroShare to set up a metadata-only record that points to a dataset that is located within another repository such as OCADS. This approach would be useful if the NEP was interested in managing multiple datasets with HydroShare, but also wanted to make sure the OCA data was findable through the dedicated OCADS repository.

While these repositories were found to be best suited to the individual test cases, they do have limitations in enabling full interoperability of data across organizations and datasets because the measurements or observations are not standardized to a common template or schema. If that level of interoperability is desired by the organization, using The Exchange Network, CUAHSI HydroClient, or another domain specific repository and conforming to the specialized data schema would be necessary.

Overall, new collaboration and online data sharing tools and approaches are available to scientists and managers to communicate results to a broad community. These approaches, such as Open Science, and the FAIR[8] data principles also offer greater opportunities to scientists to share large continuous monitoring datasets.

[7] https://essd.copernicus.org/articles/13/2777/2021/

[8] FAIR Data Principles – https://www.go-fair.org/fair-principles/

Best Practices for Preparing and Submitting Data

Choose Repository

Before submitting data to an online repository, it is important to identify the repository that is best suited for the dataset and the level of expertise of the data submitter. Understanding the repository submission process at the outset of a project can provide insight into best practices for developing metadata and structuring datasets. It is important to be familiar with the requirements from the selected data repository. Communication with the OCA community of practice, for example through the Ocean Acidification Information Exchange, is a good source of information on how to best represent data using standard names if there are minimal requirements for the data. If a standard does not exist for the type of data being submitted, it might be helpful to search for similar datasets in the repository as examples for structuring the data and metadata. For example, OCADS recommends the use of data standards and quality flag codes from earlier work of the Clivar & Carbon Hydrographic Data Office (CCHDO) and World Ocean Circulation Experiment (WOCE). OCADS is in the process of updating these standards.

Test Access and Discovery Tools

The data submitter should become familiar with the discovery tools from the selected data repository. This will be a helpful guide in making sure the data are well described and findable. Use the repository's tools to discover and evaluate similar datasets. This can help inform what information should be included to make the dataset more findable. What filters are available? Is the search conducted using open ended keywords or a curated list of selectable terms? Is the title descriptive of the dataset? Are similar datasets stored in the data repository?

Create Metadata

Independent of which data repository is ultimately chosen, creating and managing a thorough metadata document for the OCA dataset is important and can be stored as a text or PDF file. Once this information has been developed about the dataset, it can be used to add data to any repository. Depending on the requirements of the individual repository, this information can be easily transferred to a template, added to an online form, or uploaded as a file.

This file should contain all the information about the dataset to help others filter for relevance. Good metadata will enable researchers to determine if data are valuable for their needs. Most online repositories require some common descriptive metadata necessary to catalog the dataset within the repository, including project title, contributing researchers, project summary, spatial and temporal extent, parameters measured, and keywords.

To ensure data are optimized for reusability, the FAIR data principles should be applied to the metadata that describe a dataset. The FAIR principles were primarily designed to improve machine-readability (i.e., the ability use computational systems to find, access, integrate data from multiple sources (i.e., make interoperable), and reuse data with no or minimal human intervention). Even if machine-readability is not a primary consideration for the data provider, implementing FAIR data principles will improve the reusability of the data by others.

More detailed metadata to describe the parameters being measured should include the parameter names, units, calibration detail, uncertainty from sensor provider, laboratory processes, calculations, and other data specific descriptions. The repository may provide a template to enter the metadata that can be submitted with the dataset.

Basic information that should be described in a metadata document:

- Description of the program and project.
  - Include institution, title, abstract, funders, partners.
- Description of the origin of the dataset in the metadata.
  - Where were the data collected? Provide a geospatial bounding box of the study area.
  - When were the data collected? Provide a temporal range if the data are complete.
- Method of data collection.

- Indication if the data are processed or a raw dataset.
- Description of the data being collected.
    - Variables, units, vocabularies, uncertainty (from instruments).
- Description of the purpose of the dataset. Why was this dataset created? What was the goal when the dataset was created? While data may have been created for a specific purpose, describing the data may enable other purposes or applications of the data.
- Indication if there are usage limitations or considerations.
- Indication if the data can be shared openly or are they too sensitive to be shared publicly. Are the data freely available for use? For what purposes?
- Description of how to access the data (online? email? contact?). If data are available online be sure to include a get data link so the data can be easily accessed.

Submit Data and Test Access and Publication of the Dataset

Once the metadata have been created and any requirements for the dataset are met (file type or format), it can be submitted to the data repository. Depending on the repository, the data submitter may be asked to create an account or directed to send an email to an individual. Once the data are submitted, it should be verifed that the data are stored accurately, a DOI is provided (if available from the repository), and that the data can be found using the search tools.

References

Amorim, R.C., Castro, J.A., Rocha, J., and C. Ribeiro. (2014). LabTablet: semantic metadata collection on a multi-domain laboratory notebook. In: *Communications in Computer and Information Science* 478:193-205.

Duarte, C.M., Hendriks, I.E., Moore, T.S., Olsen, Y.S., *et al*. (2013). Is Ocean Acidification an Open-Ocean Syndrome? Understanding Anthropogenic Impacts on Seawater pH. *Estuaries and Coasts* 36: 221–236. https://doi.org/10.1007/s12237-013-9594-3.

Gledhill, D.K., White, M.M., Salisbury, J., Thomas, H., *et al*. (2015). Ocean and coastal acidification off New England and Nova Scotia. *Oceanography* 28(2):182–197. http://dx.doi.org/10.5670/oceanog.2015.41.

Gassett, P.R., O'Brien-Clayton, K., Bastidas, C., Rheuban, J.E., *et al*. (2021). Community Science for Coastal Acidification Monitoring and Research. *Coastal Management* 49(5):510-531. https://doi.org/10.1080/08920753.2021.1947131.

Jiang, L.-Q., O'Connor, S. A., Arzayus, K. M., and Parsons, A. R.: A metadata template for ocean acidification data, Earth Syst. SData, 7, 117–125, https://doi.org/10.5194/essd-7-117-2015, 2015.

Pimenta, A. and J. Grear. Guidelines for Measuring Changes in Seawater pH and Associated Carbonate Chemistry in Coastal Environments of the Eastern United States. U.S. EPA Office of Research and Development, Washington, DC, EPA/600/R-17/483, 2018.

Rheuban, J.E., Doney, S.C., McCorkle, D.C., and R.W. Jakuba. (2019). Quantifying the effects of nutrient enrichment and freshwater mixing on coastal ocean acidification. *Journal of Geophysical Research: Oceans* 124(12):9085– 9100. https://doi.org/10.1029/2019JC015556.

Rosenau, N.A., Galavotti, H., Yates, K.K., Bohlen, C.C., *et al*. (2021). Integrating High-Resolution Coastal Acidification Monitoring Data Across Seven United States Estuaries. *Frontiers in Marine Science* 8:679913. https://doi.org/10.3389/fmars.2021.679913.

Saba, G.K., Goldsmith, K.A., Cooley, S.R., Grosse, D., *et al*. (2019). Recommended priorities for research on ecological impacts of ocean and coastal acidification in the U.S. Mid-Atlantic. *Estuarine, Coastal and Shelf Science* 225:106188. https://doi.org/10.1016/j.ecss.2019.04.022.

Turner, J., Gassett, P., Dohrn, C., Miller, H., *et al*. (2021). Opportunities for U.S. State Governments and in-Region Partners to Address Ocean Acidification through Management and Policy Frameworks. *Coastal Management* 49(5):436-457. https://doi.org/10.1080/08920753.2021.1947126.

U.S. Environmental Protection Agency. (2021). Measuring Coastal Acidification Using *In Situ* Sensors in the National Estuary Program. Washington D.C., Document No. EPA-842-R-21001.

Washington State Blue Ribbon Panel on Ocean Acidification. (2012). *Ocean Acidification: From Knowledge to Action, Washington State's Strategic Response*. H. Adelsman and L. Whitely Binder (eds). Washington Department of Ecology, Olympia, Washington. Publication no. 12-01-015. https://cig.uw.edu/publications/ocean-acidification-from-knowledge-to-action-washington-states-strategic-response/.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., *et al*. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018. https://dx.doi.org/10.1038%2Fsdata.2016.18.

Appendix A: Glossary of Terms

Application Programming Interface (API) – Programming code that enables separate software applications to communicate even if they are written in different programming languages. APIs are typically used to programmatically retrieve or submit data from one application to another application.

CF (Climate and Forecast) Conventions – NetCDF (Network Common Data Form) Climate and Forecast metadata conventions https://cfconventions.org/.

Citation Metadata – High-level information about a data set that can be used to cite the data for use by others. Depending on the requirements of the repository, this information can be entered through a website form or submitted to data repositories as a separate file along with the file of data observations. Citation metadata generally includes the name of the study or dataset, summary of the data and parameters measured, researcher names and affiliations, dates and geospatial range of the study, and keywords that will make the data more findable.

The Darwin Core Standard (DwC) – A data standard for compiling biodiversity data from varied and variable sources. https://dwc.tdwg.org/ (see data standard below).

Data Archive – A system for the long-term storage of data. Generally, archiving is the process of moving data out of a more easily accessible system and into a long-term storage system. The terms “archive” and “repository” are used to describe systems for long-term data storage and are often interchanged. A formal data archive generally indicates there are specific procedures and rules for storage and retrieval of data. NOAA’s National Center for Environmental Information (NCEI) serves as the official Archive for data collected by NOAA scientists and NOAA funded projects and research efforts (e.g., the Integrated Ocean Observing System). Other non-NOAA data can be submitted to be archived, but must adhere to guidelines and are subject to approval. Other federal agencies may have formal data archiving policies and procedures.

Database – A structured format used to store and manage data. In a relational database structure, data are stored in tables made up of rows and columns. The relational database also contains descriptions of how the data tables relate to each other so that it can be accessed through queries. Examples of relational database software include Microsoft SQL, MySQL, PostgreSQL, and Oracle.

Data Management System – Systems used to manage, organize, and provide access to data. Data management systems can include relational databases (e.g., Oracle, SQL Server, Microsoft Access, MySQL) or flat file data storage (e.g., Microsoft Excel or .csv).

Data Repository – A data storage system containing a collection of individual datasets that have been organized in a logical manner and made accessible for use. The method of data storage used in a particular data repository can range from a structured relational database to a collection of individual files such as spreadsheets, documents, or even image or video files.

Data Schema – The structure of a database defined by how data are organized and their relationships. It is also defined by the business rules or constraints on the data. Elements include description of data fields and data types.

Data Standard – A set of specifications or rules for how data should be described and recorded. Data standards are generally developed and maintained through consensus of a group of subject matter experts. Mapping datasets to a common format appropriate to the content area of the data will enable the ability to share, combine, and better understand data from different sources. CF Conventions, Darwin Core, ISO, and WQX are examples of data standards used by the water quality community.

DOI – Digital Object Identifier, a unique string of numbers and letters used to give a unique identifier to an article, document, or dataset. Some data repositories will issue or mint a DOI for datasets contained in the repository. Others will allow users to add a DOI obtained through another provider. DOIs are increasingly being required for datasets referenced in journal publications.

ERDDAP – Open-source data management software used to set up a data repository for storing gridded and tabular scientific datasets in common file formats. Developed by NOAA's Southwest Fisheries Science Center, ERDDAP software is free and open source and can be installed and managed on a local or cloud-based server. https://coastwatch.pfeg.noaa.gov/erddap/index.html.

FAIR Data – Data that meet the principles of Findable, Accessible, Interoperable, Reusable. The GO FAIR Initiative is a stakeholder driven effort to implement FAIR principles for data. https://www.go-fair.org/fair-principles/.

FTP – File Transfer Protocol, a methodology for transferring files between computers across the internet using standard internet protocols. It is often used for files that are too large to transmit through email.

GitHub – A repository hosting platform for storing and sharing open-source software code. GitHub is also a version control system which enables collaborative development on public repositories for maintenance, upgrades, and improvements. http://github.com.

Metadata – Information that describes and is essential for understanding the data. Citation metadata often required by data repositories includes the researcher's name and affiliation, and temporal and spatial extent of where the dataset was collected. Additional descriptive metadata often required by data repositories includes a description, keywords, usage and citation information, DOI, and study design and methods to name a few. Metadata that describes the actual observations should include date and time of observation, variable names, label or description, and units at a minimum. Data repositories and data standards will have required metadata elements and example templates.

NetCDF – Network Common Data Form, a file format for storing complex scientific data that contains multiple variables, such as a data from an observing platform, satellite, or model grid (e.g., temperature, salinity, air pressure, etc.). The file format is "self-describing" which means it includes information within the file that describes the structure and layout of the file that can be interpreted by software designed to read the file type. https://www.unidata.ucar.edu/software/netcdf/.

Published data – Data that are finished, have received formal DOIs, and are discoverable and accessible by the public. Once published, data managers can no longer modify the content or metadata descriptions.

QAPP (Quality Assurance Project Plan) – A formal written document describing the detailed quality control procedures that will be used to achieve a specific project's data quality requirements. https://www.epa.gov/citizen-science/quality-assurance-handbook-and-guidance-documents-citizen-science-projects.

Vocabulary – Often referred to as controlled vocabulary, provides a consistent way to describe data using a set of standard terms (e.g., field names, units, categories). Vocabularies are developed and maintained through communities of practice. Examples include Climate and Forecast conventions and Darwin Core.

Web Service – Software that enables two machines (or applications) to send and receive data over the internet. A request is made from the client machine to the server, and information (typically data) is returned. Client and server do not have to have the same configuration to communicate when using a web service.

XML – eXtensible Markup Language is a markup language like HTML that was designed to store and transport data. What makes it extensible is flexibility within the language for a user to define custom tags that describe the data.

Appendix B: Interview Questions

Responses can be found in Appendix C.

Data Manager

- Can you describe the flow of data from sensors to internal data management system?
  - Highlight challenges and aspects that work well.
- What format are the data stored in (Excel, database)?
- Has a metadata record been developed?
- Is there a QAPP and does it address QA/QC or flags on data?
- Have the data been submitted to a data repository?
  - If so, why are they submitting data to that system, and what is the desired functionality of the data repository/ management system?
  - If not, what are the limitations?
- Who are the users/stakeholders of your data?

Data User

- How do you use OCA sensor data sets?
- How do you currently access OCA sensor data sets?
- Do you have a preferred file format (e.g., csv, MATLAB) for data?
- How important is access to metadata?
- Do you use any online data portals/repositories to access OCA other water quality data?
- What works and what doesn't work for the portal(s)?
- Do you use any online portals/repositories for submitting data?
- What works and what doesn't work for submissions?
- What functionality are you looking for in an online repository?
- What is your ideal process for accessing data?
- What is your ideal process for uploading data?
- When you download a data set do you expect to get metadata?
- What metadata do you expect to get?
- How important is a DOI in understanding origin/provenance of data (for access or for citation/publication)?
- How often do you need to contact someone associated with the data before using it? If that information was included in metadata, would it remove that burden? Does that clarification information make it back into the dataset in some way?

Appendix C. Summary of Interviews with Data Managers and Data Users

Responses to interview questions in Appendix B.

NEP	DATA COLLECTION, PROCESSING, AND STORAGE METHODS
NORTHEAST/MID-ATLANTIC	
Casco Bay - Friends of Casco Bay (FOCB)	• One continuous monitoring station off Cousin's Island, plan to deploy two more • Requires on site work to maintain, need to visit every two weeks (land access preferred) • YSI data sonde measures pH, temperature, salinity, chlorophyll, fdom, pCO_2 - happy with system, requires maintenance • Take grab samples for validation • Have existing QAPP and added to it for OCA sensors • QA/QC procedures extension of work they've always done • Like EPA document for lab, but wish there was field station QA/QC guidance • Every two weeks upload data to local database following QAPP ▪ Convert pCO_2 data, calculate TA and other OCA indicators with CO_2 ▪ Values added to Excel spreadsheet ▪ Flag QA/QC issues • Metadata are essentially ongoing log • Data used by Maine Dept of Env Protection, Univ of New Hampshire, Univ of New England, EPA and Casco Bay Estuary Program (CBEP) • Have been working with Maine DEP to get data into Water Resources Database (WRDB) ▪ Like this because other groups use the data ▪ It has been a big undertaking to convert to ME's Environmental and Geographic Analysis Database (EGAD) database first • Interested in sharing data more broadly
Casco Bay – University of New Hampshire (UNH)	• Three sensors measuring pH, salinity, temp, pCO_2 funded by Casco Bay Estuary Program (CBEP) • No telemetry, must retrieve loggers and manually download data monthly • Data processed by UNH using MATLAB into hourly data in Excel, organize into annual files • Files are stored at UNH, create annual QC file to send with data (primarily CBEP) • No telemetry, so they don't know if sensors stopped working until they are downloaded • Excel data contains basic metadata, also create annual report with additional metadata • CBEP data has not been submitted to online repository • UNH has submitted other OCA data to OCADS and SOCAT (Surface Ocean CO2 Atlas) cruises
Massachusetts Bays	• Current sensor in Duxbury, data are telemetered to UMass Boston, plans for another sensor in Barnstable • Data available on request, interest in making data accessible (considering NERACOOS) • Developing QAPP with UMass • Would be interested in IOOS Quality Assurance/Quality Control of Real-Time Oceanographic Data (QARTOD)/QA QC for OCA data • Stakeholders include shellfish industry, state agency, research
Barnegat Bay	• Three sensor systems (YSI EXO platform) • All telemeter but only one has OCA data (pH, CO_2) • YSI data goes to NJ DEP website available in real time, but no OCA data • Developing a QC process with NOAA lab for OCA • pH and CO_2 data are pulled off logger and stored as .csv files • No metadata records for OCA data • Interested in DOI, but not a requirement

NEP	DATA COLLECTION, PROCESSING, AND STORAGE METHODS
WEST COAST	
San Francisco Bay	• Data stream sent to CENCOOS and are available in ERDDAP • Following IOOS guidance working with CENCOOS (RA) to apply QARTOD and archive • For non-real-time data, periodically send to CENCOOS to be posted; not high level of QC • Maintain basic internal metadata, QA/QC • Haven't submitted to other OCA repositories, time/resource limited, hasn't been a big priority

Interviews with Data Users

DATA USER	DATA DISCOVERY AND ACCESS
SUMMARY	
NEP Program Director (Curtis Bohlen, Casco Bay Estuary Partnership)	• Prefers data in csv or Excel, but concerns with large data files and analysis software (R) • Metadata is critical and not well defined or buried in protocols ▪ Specifically, would like more clarity on how data was processed, what calculations were used, sampling design, where/when collected • Has used online repositories (NCEI, OCADS), prefers API to script data, must use different process for every repository • Data formats and vocabularies are inconsistent across data providers • DOI not critical yet, but can see growing need • CBEP has started using GitHub repository to share data and analysis from State of the Bay report
OCA Researcher (Grace Saba, PhD, Rutgers University)	• NCEI is the first stop for full datasets, generally easy to get list of matching datasets ▪ Often must search with in the results, change search parameters or contact data provider to be sure • Uses data from Barnegat Bay, but must contact directly to get data • Gets data from state repositories like NJDEP • Has gotten data from BCO DMO • Prefers NetCDF, but can work with csv or Excel • Metadata is usually lacking information about QC, often hidden in general protocols ▪ Often hard to tell if flagged data are removed or just flagged ▪ Would like to see better QC with carbonate data generally to feel confident using in analysis • DOI becoming more important for citing in reporting and with data requirements for projects
Mass Bays Estuary Partnership (Pam DiBona, Prasseda Vella)	• Using OCA data to address shellfish industry concerns, working with OCA commission • Often get data directly from organization website or regional collaborations (NE Ocean Data Portal) ▪ Data are often stale • Flexible in terms of data types, Excel is preferred if had to choose • Finds most providers have different style/narrative for metadata • Need QA/QC information to know how data was collected to not have to make assumptions • EPA requires QAPP for data providers which is helpful • Prefer data portal to discover/access data, less important to visualize and see trends over time but would find useful • DOI is not important at this stage • Often must contact data providers to understand study design and metadata • Very interested in QARTOD recommendations for OCA data

Appendix D. Data Repository Submission Detail

Disclaimer: This section was accurate at the time of publication. The process for submitting data to the repositories is subject to change or updates. Please visit the repository website for updated information or guidance before submitting data.

NCEI OCADS

Submitting Data

The OCADS data portal provides support and guidance for preparing data for submission. The recommended approach is to prepare metadata following a template provided (Metadata submission form). This template is an Excel file with metadata elements (both required and recommended) for the user to provide input specific to the dataset. There are supporting documents for help on metadata element names (Instruction file). Of the 221 elements, 36 are required and several are repeat blocks for multiple variables. The global variables include basic program level detail (PI name and institution, title of project, abstract, author list). The main component of the metadata is the "variable metadata" section. Required elements include variable abbreviation, unit, uncertainty. Specific variable blocks for OCA variables (DIC, TA, pH, pCO_2A, pCO_2D) are provided to provide additional detail where appropriate (e.g., for pH, temperature of measurement, temperature recorded, uncertainty). Other values (e.g., date, longitude, latitude, depth, bottle number, QA flags, standard deviation) that are not considered independent don't require separate sections. It is recommended to avoid special characters (degree symbol, sigma, etc.). Additional information is available through a 2015 paper that describes a metadata template for ocean acidification data (Jiang et. al., 2015). Work is currently underway to update these standards.

Guidance for preparing data files is available through additional templates based on the type of data. These include Underway, Profile data (e.g., CTD, discrete bottle water samples), and Mooring data (e.g., buoys). The templates provide recommendations for column header names and an example data file for each category of data. No specific file format is required, but ASCII (.csv) or NetCDF is recommended, and proprietary file types are to be avoided if possible. Using CCHDO parameters and WOCE quality control flags is also recommended whenever possible.

Once the data templates are complete, the metadata and data files are then uploaded through the NOAA PMEL Scientific Data Information System (SDIS) (https://data.pmel.noaa.gov/sdig/oap/Dashboard/OAPUploadDashboard.html). A video tutorial is available on the website. If the data files are too large (>20 MB), NCEI will provide an alternative method for transferring data.

1. Underway Column header names description (data file (csv) example).
2. Profile data (e.g. CTD, discrete bottle water samples) Column header names description (data file (csv) example).
3. Mooring data (e.g. buoys) Column header names description (data file (csv) example).

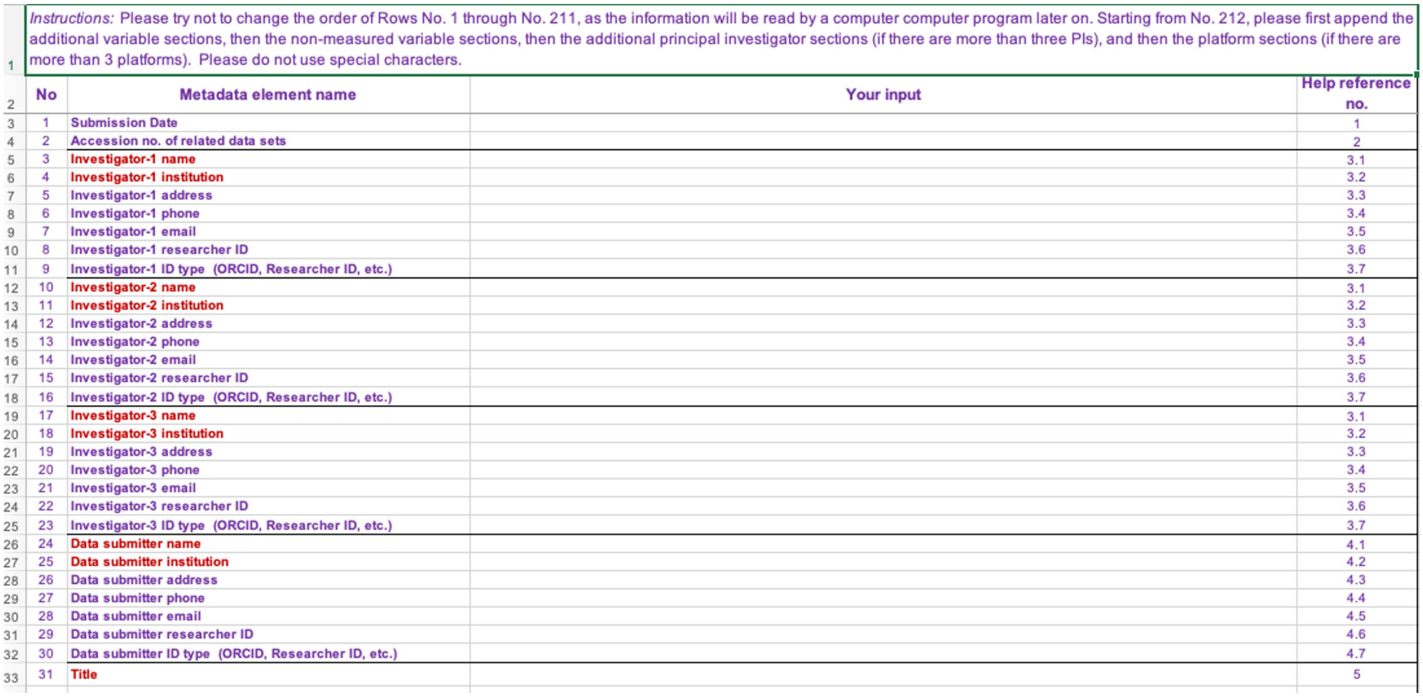
While the data are mapped to a template, the search and discovery tools only search metadata and data are downloaded in their original file format.

Acquiring Data from OCADS

OCADS provides a data access portal (https://www.ncei.noaa.gov/access/oads/) to discover and acquire datasets. Several filters are available to narrow down the selection of data. They include filters for: "Core variables" (e.g., TA, DIC, pH, CO_2, temp, salinity, nutrient, etc.); "Other variables" (an uncurated list of user submitted variables); "Observation category" (e.g. surface underway, profile, time series, model output, benthic FOCE); "Additional terms" (free text keyword search) "Observation date" (start/end date), and a map where users can enter bounding box coordinates to filter geospatially.

The results are sorted by most recently submitted data, but no additional sorting options are given at this time. The results include the title, first two lines of abstract, a thumbnail image of sample area and links to NCEI metadata and Project metadata. The number of matching results is indicated with a link to further refine the search. The full results can be accessed through web services (RSS, ATOM, KML, JSON, CSV).

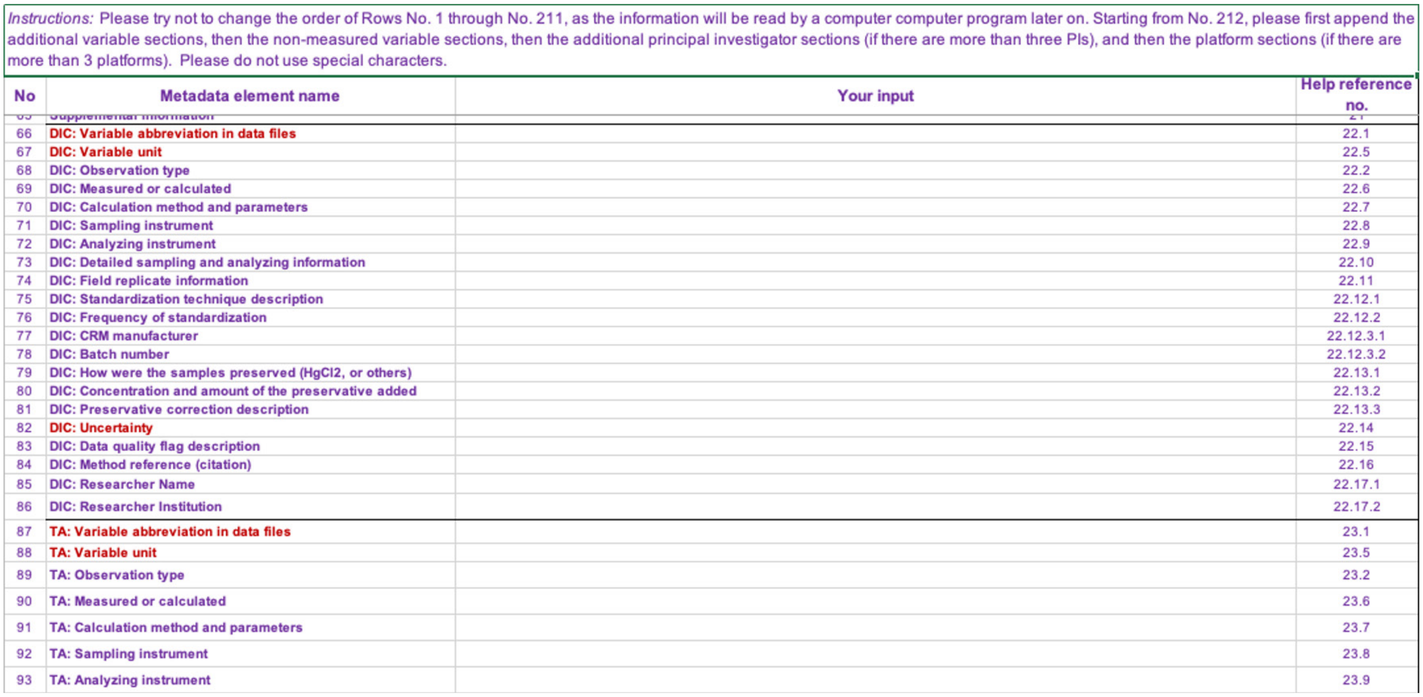
The metadata page includes the full description of the dataset, including contact information, citation information, DOI, usage constraints, etc. From the metadata page, data can be downloaded via HTTPs or FTP.



1 *Instructions:* Please try not to change the order of Rows No. 1 through No. 211, as the information will be read by a computer computer program later on. Starting from No. 212, please first append the additional variable sections, then the non-measured variable sections, then the additional principal investigator sections (if there are more than three PIs), and then the platform sections (if there are more than 3 platforms). Please do not use special characters.

2	No	Metadata element name	Your input	Help reference no.
3	1	Submission Date		1
4	2	Accession no. of related data sets		2
5	3	Investigator-1 name		3.1
6	4	Investigator-1 institution		3.2
7	5	Investigator-1 address		3.3
8	6	Investigator-1 phone		3.4
9	7	Investigator-1 email		3.5
10	8	Investigator-1 researcher ID		3.6
11	9	Investigator-1 ID type (ORCID, Researcher ID, etc.)		3.7
12	10	Investigator-2 name		3.1
13	11	Investigator-2 institution		3.2
14	12	Investigator-2 address		3.3
15	13	Investigator-2 phone		3.4
16	14	Investigator-2 email		3.5
17	15	Investigator-2 researcher ID		3.6
18	16	Investigator-2 ID type (ORCID, Researcher ID, etc.)		3.7
19	17	Investigator-3 name		3.1
20	18	Investigator-3 institution		3.2
21	19	Investigator-3 address		3.3
22	20	Investigator-3 phone		3.4
23	21	Investigator-3 email		3.5
24	22	Investigator-3 researcher ID		3.6
25	23	Investigator-3 ID type (ORCID, Researcher ID, etc.)		3.7
26	24	Data submitter name		4.1
27	25	Data submitter institution		4.2
28	26	Data submitter address		4.3
29	27	Data submitter phone		4.4
30	28	Data submitter email		4.5
31	29	Data submitter researcher ID		4.6
32	30	Data submitter ID type (ORCID, Researcher ID, etc.)		4.7
33	31	Title		5

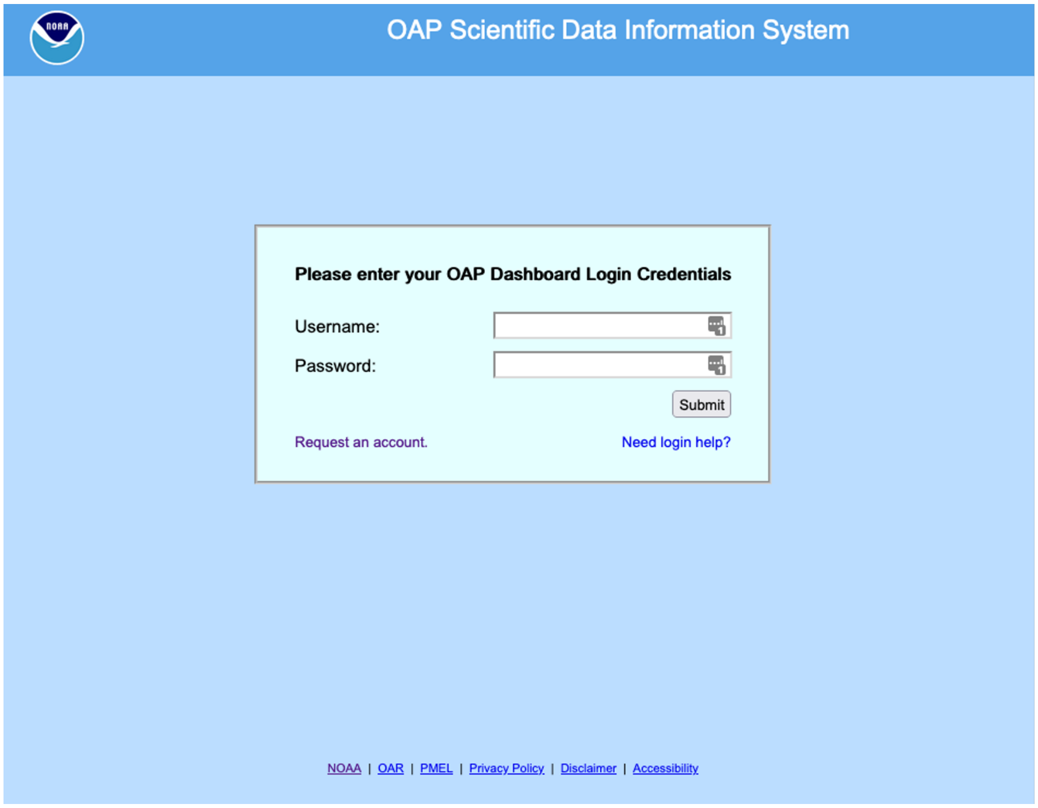
Figure 3 - screen capture of OCADS metadata template showing required fields (in red)



Instructions: Please try not to change the order of Rows No. 1 through No. 211, as the information will be read by a computer computer program later on. Starting from No. 212, please first append the additional variable sections, then the non-measured variable sections, then the additional principal investigator sections (if there are more than three PIs), and then the platform sections (if there are more than 3 platforms). Please do not use special characters.

No	Metadata element name	Your input	Help reference no.
65	Supplemental information		21
66	DIC: Variable abbreviation in data files		22.1
67	DIC: Variable unit		22.5
68	DIC: Observation type		22.2
69	DIC: Measured or calculated		22.6
70	DIC: Calculation method and parameters		22.7
71	DIC: Sampling instrument		22.8
72	DIC: Analyzing instrument		22.9
73	DIC: Detailed sampling and analyzing information		22.10
74	DIC: Field replicate information		22.11
75	DIC: Standardization technique description		22.12.1
76	DIC: Frequency of standardization		22.12.2
77	DIC: CRM manufacturer		22.12.3.1
78	DIC: Batch number		22.12.3.2
79	DIC: How were the samples preserved (HgCl2, or others)		22.13.1
80	DIC: Concentration and amount of the preservative added		22.13.2
81	DIC: Preservative correction description		22.13.3
82	DIC: Uncertainty		22.14
83	DIC: Data quality flag description		22.15
84	DIC: Method reference (citation)		22.16
85	DIC: Researcher Name		22.17.1
86	DIC: Researcher Institution		22.17.2
87	TA: Variable abbreviation in data files		23.1
88	TA: Variable unit		23.5
89	TA: Observation type		23.2
90	TA: Measured or calculated		23.6
91	TA: Calculation method and parameters		23.7
92	TA: Sampling instrument		23.8
93	TA: Analyzing instrument		23.9

Figure 4 – screen capture of OCADS metadata template showing additional required metadata for describing data parameters (in red)



OAP Scientific Data Information System

Please enter your OAP Dashboard Login Credentials

Username:

Password:

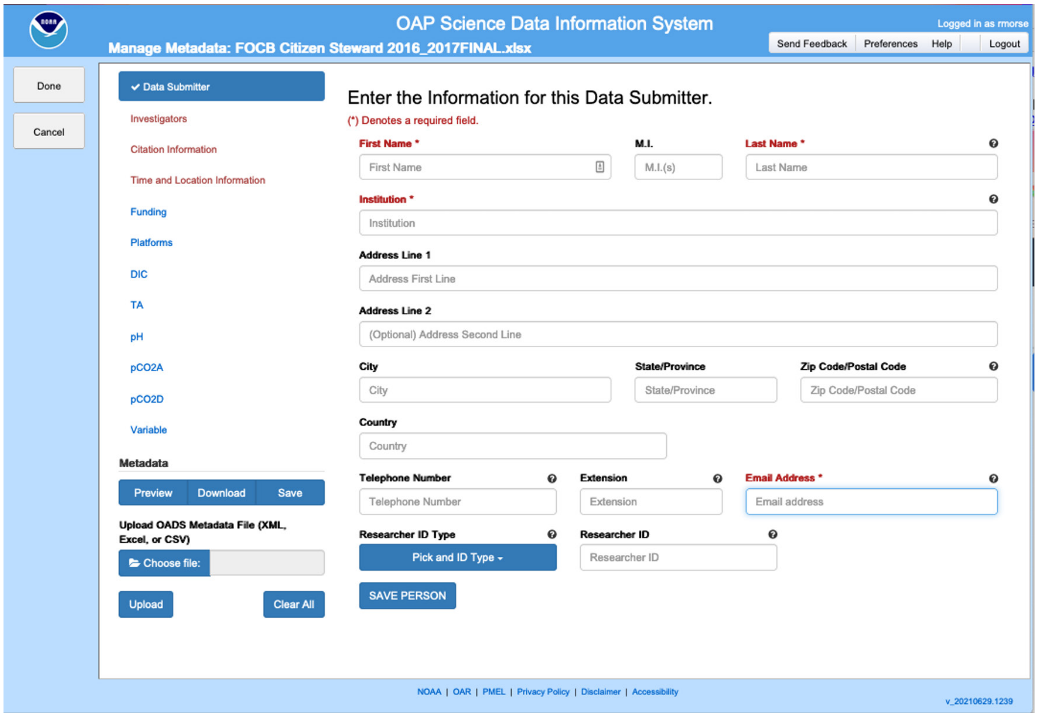
Submit

Request an account.

Need login help?

NOAA | OAR | PMEL | Privacy Policy | Disclaimer | Accessibility

Figure 5 – screen capture of Ocean Acidification Program Scientific Data Information System (OAP SDIS) login page



OAP Science Data Information System

Manage Metadata: FOCB Citizen Steward 2016_2017FINAL.xlsx

Logged in as rmorse

Send Feedback | Preferences | Help | Logout

Done

Cancel

✓ Data Submitter

Investigators

Citation Information

Time and Location Information

Funding

Platforms

DIC

TA

pH

pCO2A

pCO2D

Variable

Metadata

Preview | Download | Save

Upload OADS Metadata File (XML, Excel, or CSV)

Choose file:

Upload

Clear All

Enter the Information for this Data Submitter.

(*) Denotes a required field.

First Name * — First Name

M.I. — M.I.(s)

Last Name * — Last Name

Institution * — Institution

Address Line 1 — Address First Line

Address Line 2 — (Optional) Address Second Line

City — City

State/Province — State/Province

Zip Code/Postal Code — Zip Code/Postal Code

Country — Country

Telephone Number — Telephone Number

Extension — Extension

Email Address * — Email address

Researcher ID Type — Pick and ID Type

Researcher ID — Researcher ID

SAVE PERSON

NOAA | OAR | PMEL | Privacy Policy | Disclaimer | Accessibility

v_20210629.1239

Figure 6 – screen capture of OAP SDIS metadata entry form

CUAHSI HydroShare

Submitting Data

Data submitters can publish and share datasets, manage access to shared content for collaborators, obtain a digital object identifier (DOI) that can be used in publications and citations, and aggregate resources into collections. Once in HydroShare, data resources can be readily discovered and acquired. HydroShare requires that data submitters create an account before uploading data; this is done by simply providing an email address and password, and validating the link.

Support is available through a collection of FAQs (https://help.hydroshare.org/) as well as a tutorial to describe the process for uploading data. The guidance recommends organizing data in advance of submission, especially where there are multiple files to be submitted. Recommendations include taking into consideration how potential end-users would want to interact with the data resources and the structure that will enable access, interpretation, and reuse.

Each data submission is considered a "resource" which is defined as the data and associated metadata for a single unit of digital content. Comparable terms used by other systems include "package" or "dataset." Within a resource on HydroShare, multiple files of different types can be grouped within one resource and examples are provided on how one might structure individual data resources. Another level of organization called a "collection" is available, where multiple, individual data resources can be grouped together under a common theme such as an event (e.g., Hurricane Harvey) or a network of groups monitoring similar data.

A template is provided to make it easier to add metadata for the resource. There are only a few required fields for the metadata: title, abstract, and keywords. Recommended fields include temporal and spatial coverage of the resource. Additional metadata fields are available to add references, sources, related resources, credits (funding agencies), and contributors. Best practices for generating the content are provided as well as guidance for naming files.

A guide is provided for data authors and publishers to provide guidance for structuring the submitted data files and metadata. The guidance is to use file formats that are open and documented standards with widespread usage in the community of practice. Tabular data should be submitted as .csv rather than Excel, with additional .csv files instead of workbooks. Each column in a table should have a detailed description in the abstract, in a README file, or in a descriptive header within the file that contains the data table. This includes the full name of the variable, a description of what it represents, units, and how it was obtained for each column. A workflow guideline is provided for submitting data that are in a complex Excel workbook with multiple tabs for raw data, finished products, and intermediate steps. Specific guidance for data and common vocabulary or standard names is not provided.

Once the data resource is added, it can be made private or shared in three states - discoverable, public, or published. These are described as follows:

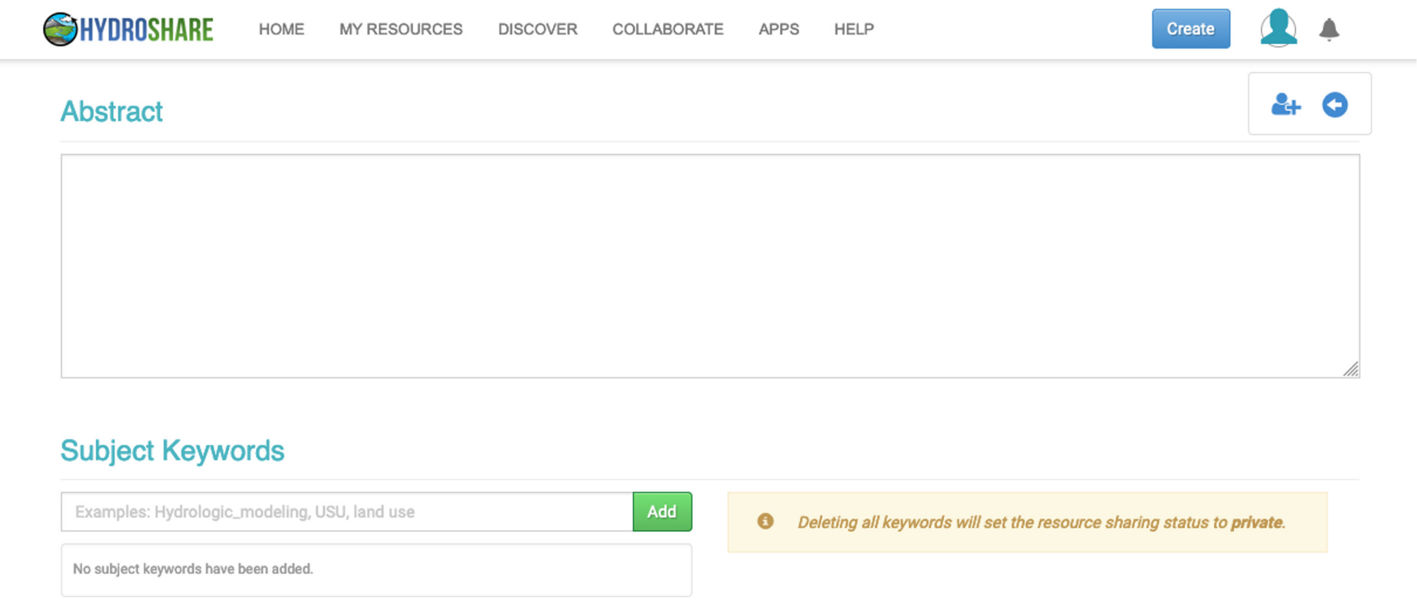
- **Discoverable** – A resource that can be discovered by anyone in HydroShare, but only users with permission can access the content files.
- **Public** – A resource that can be discovered by anyone, and anyone can access the files. The resource may not be final, may be subject to change, could be deleted, and is not considered to be published.
- **Published** – A resource that is finished and is formally published. Published resources receive formal DOIs and, once published, users can no longer modify their content or metadata descriptions. Published resources can be discovered and accessed by the public.

Upon first submitting a resource and before publishing, data submitters can obtain the actual citation and DOI that will be finalized once the data are published. This facilitates providing the information to a journal that may have a deadline associated with publication.

There are no specific data structure requirements for data resources which makes the platform very flexible but provides fewer supports for data providers. This is meant to allow flexibility for data that needs to conform to a particular data model or resource type, and data providers can define specific file formats, syntax, and file hierarchies. The idea is that the data detail is well described by resource-level metadata to include information for all resources and extensions for each resource type.

Discovering Data

HydroShare has a data discovery tool (https://www.hydroshare.org/search/) to find public and discoverable resources. Filters include selecting a temporal range (using a date selection) and geospatial range (by selecting a spatial boundary from a map). Additional filters are available for categories such as author, subject, resource type, and status (public or published). The filter terms are generated from the submitted data resources and are uncurated. A text field is provided for searching via keywords.



Abstract

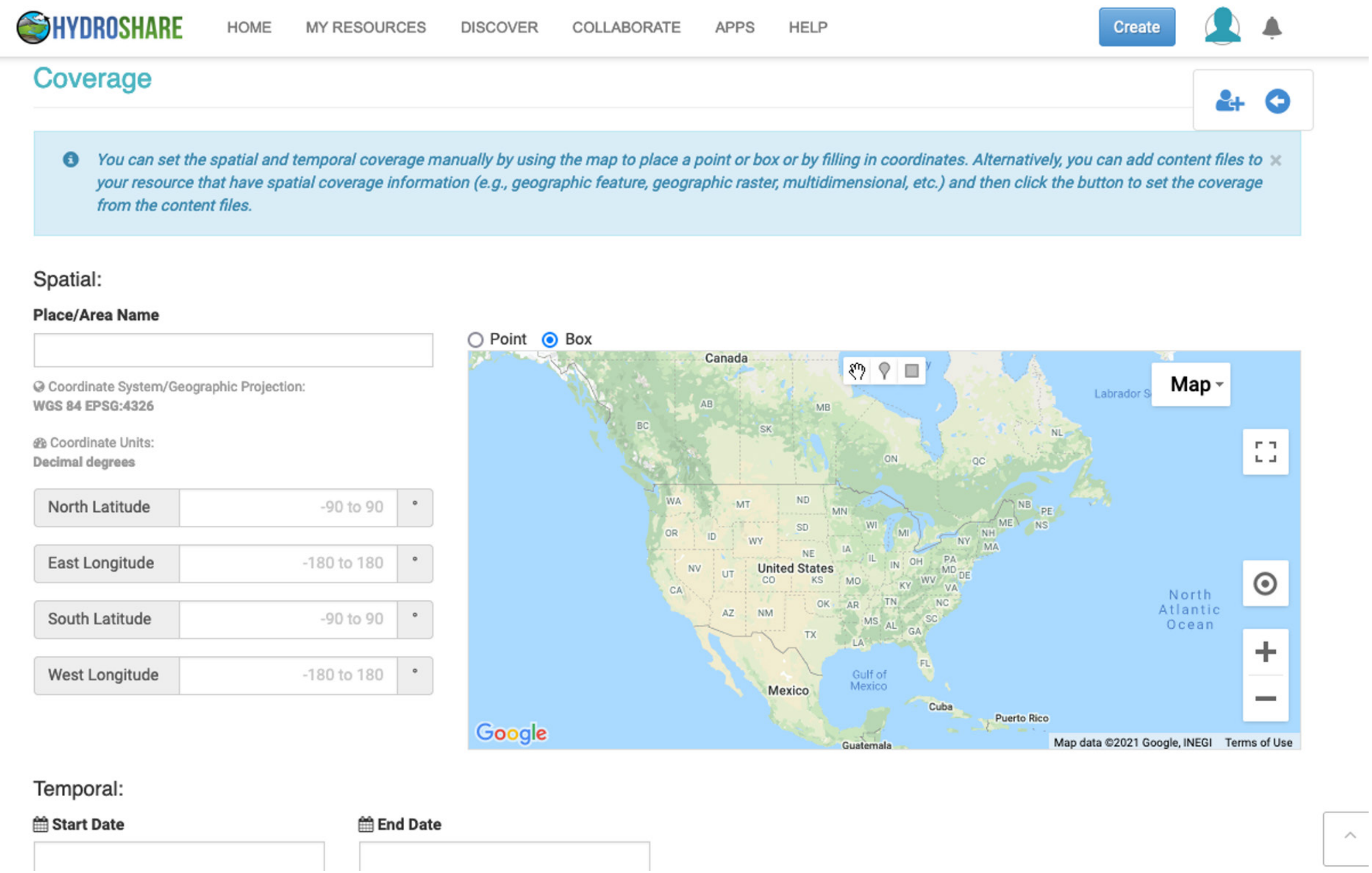
Subject Keywords

Examples: Hydrologic_modeling, USU, land use

No subject keywords have been added.

*Deleting all keywords will set the resource sharing status to **private**.*

Figure 7 - screen capture of HydroShare metadata submission form showing abstract and keyword fields



Coverage

You can set the spatial and temporal coverage manually by using the map to place a point or box or by filling in coordinates. Alternatively, you can add content files to your resource that have spatial coverage information (e.g., geographic feature, geographic raster, multidimensional, etc.) and then click the button to set the coverage from the content files.

Spatial:

Place/Area Name

Coordinate System/Geographic Projection:
WGS 84 EPSG:4326

Coordinate Units:
Decimal degrees

North Latitude	-90 to 90	°
East Longitude	-180 to 180	°
South Latitude	-90 to 90	°
West Longitude	-180 to 180	°

○ Point ● Box

Temporal:

Start Date

End Date

Figure 8 - screen capture from HydroShare showing geospatial and temporal coverage fields

CUAHSI HydroServer/HIS

The original data repository developed by CUAHSI is known as HydroServer (https://hydroserver.cuahsi.org/) and is part of the Hydrological Information System. This data repository was a precursor to HydroShare and consists of a relational database and data services. HydroServer is a more robust data integration effort and is designed primarily for time series data. The platform provides detailed templates in Excel for users to format their individual dataset. The data must be mapped to the template, and there are six required tables (general metadata) and seven optional tables. Once data are submitted to HydroServer, they are integrated into a relational database at CUAHSI, and the time series data can be discovered through a web application called HydroClient (https://data.cuahsi.org/). While this data repository requires more work on the part of the data submitter to prepare and submit, the data values become queryable and can be manipulated and visualized as basic plots through available tools and applications. This platform was designed to support hydrological data and may be more challenging to map OCA data to the provided templates. An interest in integrating a wider variety of data (models, geospatial) led to the development of HydroShare, which is more of a data catalog with no specific requirements on data type or format.

Environmental Information Exchange Network (EN)

Submitting Data

https://acwi.gov/monitoring/pubs/misc/publishing_your_data_wqp.pdf

Data are mapped to one of several available data schemas, called Data Exchanges, and submitted to the network (e.g., WQX for water data). Data are stored and managed through a centralized database at EPA (CDX/STORET). The EN is primarily for regulatory data from states, tribes, and some citizen monitoring data. Submitting to the EN typically requires working with a state agency to submit data to EPA. The WQX is the most appropriate standard for OCA data, but has limited capacity to handle continuous monitoring data.

Accessing Data

Data submitted to the Exchange Network WQX can be access through the National Water Quality Monitoring Council Water Quality Portal:

https://www.waterqualitydata.us/

IOOS RA ERDDAP

Submitting Data

Each IOOS Regional Association manages data independently for their region and most of the eleven RAs have an ERDDAP server for distribution and access to data. If an NEP is interested in working with the local RA to put OCA data into ERDDAP, the first step would be to make contact to discuss the process for submitting data, which generally involves describing the dataset (parameters, units) and providing basic metadata about the dataset.

Accessing Data

Once datasets are integrated into the RA ERDDAP, they can be accessed directly from the dedicated URL. For example, the San Francisco Estuary Partnership OCA data can be accessed through the Central and Northern California Ocean Observing System (CENCOOS) ERDDAP:

https://erddap.cencoos.org/erddap/tabledap/tiburon-co2.html

General Comments

Feedback on the data access portals was collected from the interviews. Common issues included difficulty navigating and filtering results. The titles aren't always descriptive of the dataset and requires a bit of digging into the data. It was also easy to limit results by over selecting terms and keywords.

The query results are only as good as the data entered by the data provider, reinforcing the need to develop and provide good, descriptive metadata.