International Year of the Salmon Data Mobilization and Communication Plan

Hakai Institute

Tim van der Stap, Brett Johnson

1713 Hyacinthe Bay Road, Heriot Bay, BC, Canada

Last Updated: 2021-04-22 11:10:51

Executive Summary

Effective data management for multi-disciplinary research requires both that data are accessible to current and future users, and that there is consistency in collection protocols, metadata standards, data assembly, quality control and publication. This Data Mobilization and Communication plan describes how IYS Data Scientists at the Hakai Institute will fulfill these requirements with the 'International Year of the Salmon' fisheries and oceanographic research expedition data. This approach can be applied to each data type collected on the expedition, and describes methods for data standardization, mobilization, and integration under the Global Ocean Observing System (GOOS) framework which identifies Essential Ocean Variables. The various data repositories, software, and technologies that support this standardization process are described in sufficient detail for an introductory overview. To achieve effective data mobilization, data should be findable, accessible, interoperable, and reusable (FAIR). Where community practice has adopted existing data mobilization paradigms, a federated approach wherein existing standards, repositories, and technologies are integrated or linked together will be used. Adopting international standards including the Darwin Core Archive (DwC-A), ISO 19115 metadata records, controlled vocabulary servers, and applying them to Essential Ocean Variables identified by GOOS will ensure a multilateral approach to the standardization of salmon ocean ecology data. Lastly, a communication strategy is proposed describing how to effectively communicate this data mobilization and standardization approach throughout the IYS project with the aim to improve understanding, collaboration, and engagement with the IYS Expedition Scientists involved.

List of Acronyms

Below is a list of the acronyms mentioned throughout this Data Mobilization and Communication plan and their description.

- BODC – British Oceanographic Data Center
- CF – Climate and Forecast
- CKAN – Comprehensive Knowledge Archive Network

- DwC(-A) – Darwin Core Archive
- ERDDAP – Environmental Research Division Data Access Protocol
- EML – Ecological Markup Language
- CIOOS – Canadian Integrated Ocean Observing System
- FAIR – Findable, Accessible, Interoperable, Reusable
- FRDR – Federated Research Data Repository
- GOOS – Global Ocean Observing System
- IOC – International Oceanographic Commission
- IPT – Integrated Publishing Toolkit
- IYS – International Year of the Salmon
- MBON – Marine Biodiversity Observation Network
- NERC – Natural Environment Research Council
- NPAFC – North Pacific Anadromous Fish Commission
- OBIS – Ocean Biodiversity Information System
- TDWG – Biodiversity Information Standards (formerly: The International Working Group on Taxonomic Databases)
- UNESCO – United Nations Educational, Scientific and Cultural Organization
- WoRMS – World Register of Marine Species

1. Introduction

The North Pacific Anadromous Fish Commission (NPAFC) is implementing a five-year International Year of the Salmon (IYS) collaborative project through 2022 to set the conditions for the resilience of salmon and people in a rapidly changing world. One of the thematic goals of the IYS is that ‘information systems that house and mobilize data about salmon and their environment are made freely available’, citing that rapid access to standardized data is one of the largest barriers to salmon research. Therefore, this Data Mobilization and Communication plan aims to describe the methods used to standardize, mobilize, and integrate data collected during the IYS research cruises to bridge gaps between the scientific domains of salmon ocean ecology, and facilitate data integration and reuse. Through active collaboration and engagement with IYS Expedition Scientists throughout the data mobilization process, we strive to share the knowledge and methods to improve the quality of the (meta)data to deepen the impact of research through transdisciplinary data integration. Developing, clearly communicating, and training IYS scientists in a workflow for how salmon ocean ecology data can be standardized, mobilized, and integrated with the Global Ocean Observing System (GOOS) framework should not only facilitate FAIR data production and management for the IYS, but also result in a lasting legacy of improved awareness of data mobilization practices for salmon ocean ecology generally.

The GOOS is a program of the International Oceanographic Commission (IOC) under UNESCO but relies on partnerships with various organizations worldwide. Some of the core objectives of the GOOS are, among others, to set the global standards and best practices for ocean-related data collection, curation (standardization), and mobilization. Expert agencies are partnered with the GOOS in biological data, including the Ocean Biodiversity Information System (OBIS) and the Marine Biodiversity Observation Network (MBON). These organizations promote and develop the use of standardized terminology ("controlled vocabularies"), metadata, and data structures. By adopting international standards, data may become interoperable with other large data sets and thus, more likely to be reused. The GOOS focuses on distinct Essential Ocean Variables (EOVs) within three domains: Physical & Climate, Biogeochemistry, and Biology & Ecosystems. The EOVs, identified by the GOOS Expert Panels, are data elements that are determined to have high impact and feasibility, and assessments of these ensure that the best, most cost-effective plan is adopted across platforms to obtain data.

The approach described in this plan relates each scientific discipline within salmon ocean ecology to one of the three domains identified in the GOOS framework and each variable collected by the expeditions to an EOV. Controlled vocabulary terms hosted on GOOS-affiliated 'vocabulary servers' are then applied to the methods, units, values, sampling platforms, and sampling protocols used in salmon ocean ecology. While standardized variable names facilitates data interoperability, formatting metadata to international standards enables data discoverability. Using the ISO 19115 geospatial metadata standard for all data sets, as well as the Ecological Markup Language (EML) for biological data enables data to be interpreted by computers and opens up data sets to a whole world of possibilities for computer-aided manipulation, distribution, integration, and long-term reuse. This document is supplemental to the International Year of the Salmon Data Mobilization Strategic Recommendations report and addresses in more detail the operational component of the standardization workflow. This document is created for internal use between the IYS Data Scientists (Hakai Institute), the IYS Secretariat, the NPAFC Study Group on High Seas Data Standardization / Mobilization ('Study Group'), and the IYS Cruise Planning Team, consisting of the National Lead Scientists, the Expedition Chief Scientist and a few others. Once finalized, it can be distributed to Principal Investigators (PIs) and IYS Expedition Scientists (see section 3) to improve understanding of the data mobilization process, the communication workflow, and best practices adopted. The document outlines the rationale behind choosing specific platforms, digital repositories and frameworks, and their associated data principles, and how the standardization process will help integrate salmon ocean ecology in a larger framework. This approach must consider ethical, legal, and scientific merits of sharing knowledge often complicated by differences in cultural systems of knowing. As the IYS is an international collaboration between Canada, Japan, Korea, Russia and the United States, it is imperative that the roles and responsibilities, and the licensing are clearly outlined and agreed upon in writing in advance of the expeditions.

1.1 FAIR Data Principles

Successful science depends on how standardized, integrated and accessible data are. Therefore, it is important that the data are open source and follow the FAIR data principles: the data should be Findable, Accessible, Interoperable, and Reusable (Tanhua et al. 2019).

- **F**indable: Data and supplemental materials need to have sufficiently rich metadata and a unique and persistent identifier.

- **A**ccessible: Metadata and data are understandable to humans and machines. Data are deposited in a trusted, secure repository.
- **I**nteroperable: Metadata use a formal, accessible, shared and broadly applicable language for knowledge representation.
- **R**eusable: Data and collections have clear usage licenses and provide accurate information on provenance.

1.2. Objectives

The objectives of this Data Mobilization and Communication plan are three-fold:

1. Describe the IYS data mobilization cyber infrastructure, and how it complies with the FAIR data principles (section 2.1);
2. Provide an overview of the workflow for data standardization, mobilization and integration (section 2.2);
3. Propose an effective data mobilization communication strategy for the 2022 multi-vessel Pan-Pacific Winter High Seas Expedition (section 3).

2. Methods

The Data Mobilization and Communication plan is applied using the principles of FAIR data, and protocols and standards for archiving and providing open access to data through the various cyber-infrastructure components affiliated with the GOOS, or widely adopted by the GOOS community.

In this document, three stages of data quality are identified for the purposes of the IYS Data Mobilization workflow to clarify where each stage of data will be stored and managed. The location of each stage of data may vary depending on each Research Area team's Data Management Plan. The Standardized Data, however, will all be stored on various national or regional ERDDAP servers or the Ocean Biodiversity Information System's (OBIS) servers and federated via the IYS Metadata Catalogue which acts as a single point of entry to access all IYS data across a distributed network of data stores.

Here, we define Raw, Processed, and Standardized Data as follows:

Raw Data: The data that are in the custom format of how they were collected and entered, or the format that is required by the scientists' home institution.

Processed Data: Data that results from data providers renaming, re-formatting, and restructuring various elements of the Raw Data to conform with the Global Ocean Observing System (GOOS) and IYS data conventions including the use of controlled vocabulary terms, standard date and location data formats, and documentation describing data collection methods, and data processing steps. These requirements will be further defined in the Practical Data Standardization Guide.

Standardized Data: Data that results from IYS Data Scientists quality checking and further restructuring of the Processed Data that will be published to OBIS or ERDDAP.

2.1 Cyberinfrastructure

Findability: Metadata catalogues host records that conform to a standard metadata profile. Using a standard metadata profile permits machine-readability for automated data discovery by assigning a globally unique and persistent identifier to the metadata record and registering that identifier in an indexed and searchable resource. The IYS Data Mobilization plan primarily depends on CKAN to store metadata records and serve as a data portal that provides not only consistent metadata but also persistent links to where data are stored. In CKAN the ISO 19115 Geographic Metadata Standard is used, a widely adopted standard for data discovery developed in by NASA's Earth Science Division. This ensures that various indexing services (such as Google) can find IYS data. All data sets receive a CKAN metadata record, while biological data hosted on the OBIS will also receive an Ecological Markup Language (EML) metadata record. For biological data, the IYS CKAN metadata record will link to the more detailed EML record (see Figure 1). OBIS EML uses a much richer collection of metadata specific to biological and ecological data and enhances findability.

Accessibility: After users have found a record of the data, they must also be able to openly access and download the data. Biological data from the IYS will be openly accessible via OBIS in a standard set of tables in the Darwin Core Archive (DwC-A) format, an .xml metadata file describing data set structure and an eml.xml metadata file describing ecological aspects of the data. These data can be downloaded as the individual data sets that were uploaded, or as part of a much larger automatically integrated data set based on data queries submitted by OBIS users. Biogeochemical and physical data do not have a globally adopted standard data set structure that permit automatic integration with other data sets. Therefore, these data are uploaded to an ERDDAP data server which is a powerful open data access platform that serves individual data sets in a consistent way, translates between common file formats, allows users to query datasets based on variables in the data, date/times, locations, institutions, and can provide subsets of specific data sets. In ERDDAP, metadata are available in multiple formats including an ISO 19115 format that describe both the dataset, and the variables in the dataset.

Interoperability: Data often need to be integrated with other similar data. This process can be difficult, error-prone and time-consuming if standardized and formal language is not used to describe data, metadata, or knowledge representation systems. An effective tool to ensure data interoperability is the use of controlled vocabularies. The IYS Data Mobilization plan depends on three controlled vocabularies: 1) the British Oceanographic Data Center's Natural Environment Research Council (BODC NERC) Vocabulary Server which broadly houses many terms for both biological data, physical and chemical data, as well as sampling instrument platforms; 2) the Climate and Forecast (CF) Conventions, which apply mostly to chemical, physical, and atmospheric variables; and 3) the World Registry of Marine Species (WoRMS) which is exclusively used for taxonomic purposes. Each of these controlled vocabularies has a very active community for maintaining and developing terms, and explicit governance structures which have allowed wide adoption among research communities.

Reusability: A key component of being able to reuse data is the inclusion of an open access data usage license. To that end, we are recommending that every dataset be licensed under the Creative Commons Attribution 4 License (CC BY 4). This license is most in line with how academic data are often expected to be attributed and cited to the original author, while permitting reuse. Standardized data structures are another method to make it easy to reuse and interoperate data. For biological data we recommend the DwC-A format and more specifically the OBIS-ENV-DATA format which is a specific implementation of the DwC-A schema. The Darwin Core[1] (DwC) is the body of standards for biodiversity used in OBIS and elsewhere. DwC provides terms and vocabularies used to format data to an international standard. Darwin Core terms and vocabularies are maintained by TDWG (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). Aside from hosting biodiversity data, OBIS can also host habitat, environmental, biotic and biometric data via the OBIS ENV-DATA format.

[1] https://obis.org/manual/darwincore/

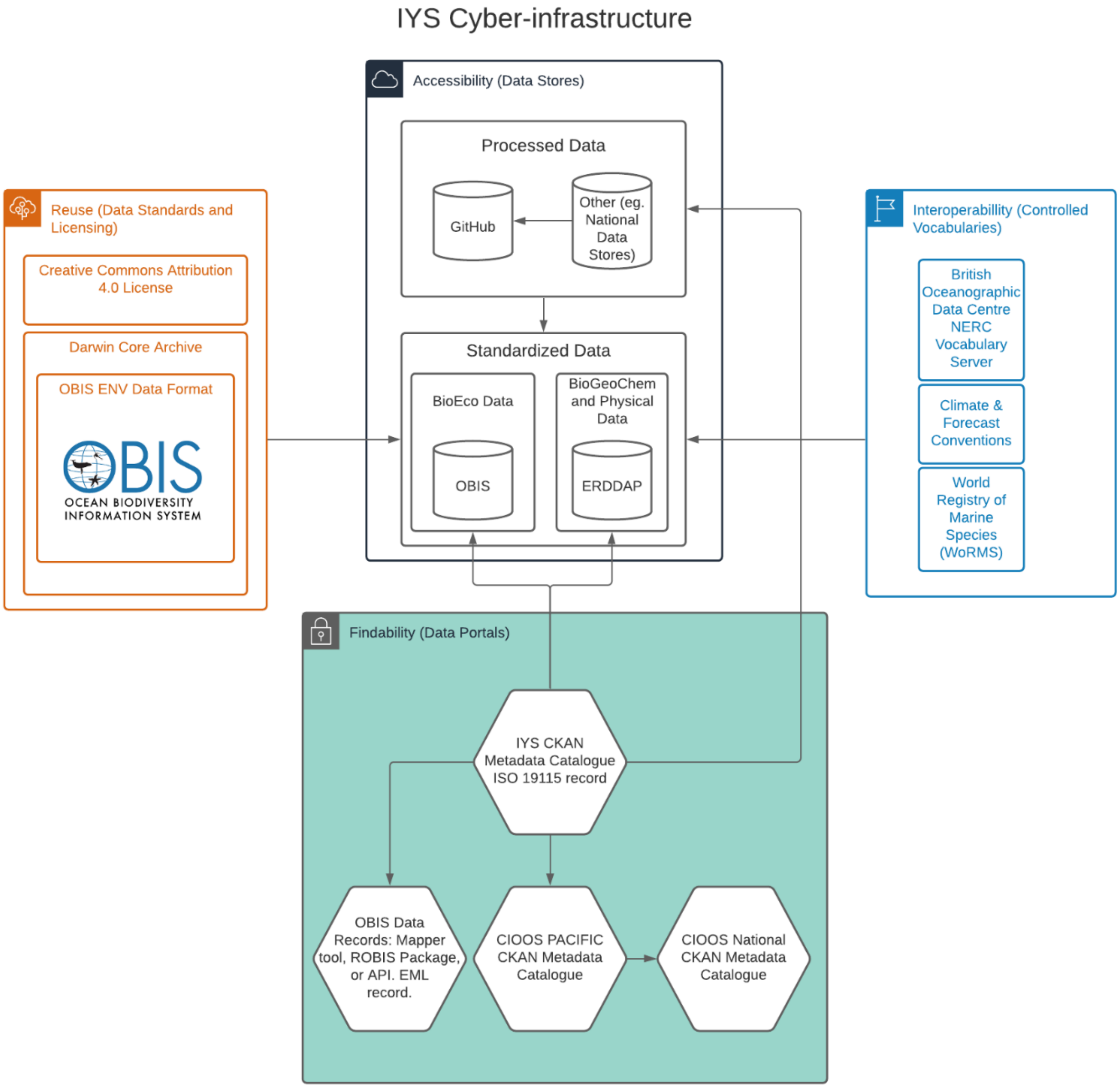
IYS Cyber-infrastructure
Accessibility (Data Stores)
Processed Data
GitHub
Other (eg. National Data Stores)
Standardized Data
BioEco Data
OBIS
BioGeoChem and Physical Data
ERDDAP
Reuse (Data Standards and Licensing)
Creative Commons Attribution 4.0 License
Darwin Core Archive
OBIS ENV Data Format
OBIS
OCEAN BIODIVERSITY INFORMATION SYSTEM
Interoperabillity (Controlled Vocabularies)
British Oceanographic Data Centre NERC Vocabulary Server
Climate & Forecast Conventions
World Registry of Marine Species (WoRMS)
Findability (Data Portals)
IYS CKAN Metadata Catalogue ISO 19115 record
OBIS Data Records: Mapper tool, ROBIS Package, or API. EML record.
CIOOS PACIFIC CKAN Metadata Catalogue
CIOOS National CKAN Metadata Catalogue

Figure 1: Elements of the International Year of the Salmon data mobilization cyber-infrastructure as they relate to the FAIR data principles.

2.2. Data Standardization Workflow

2.2.1 Data Management Plans

Ahead of the 2022 multi-vessel Pan-Pacific Winter High Seas Expedition, IYS Expedition Scientists will be asked to fill out a Data Management Plan (DMP) specific to each Research Area. The DMP will include different sections which will have to be filled in separately for each vessel. This DMP will identify similarities and discrepancies in data collection methods between the vessels and improve understanding of the format of data collected, necessary documentation, and storage and sharing protocols. It will help identify every data element, method, sampling platform, and variable that IYS Expedition Scientists plan to collect. Additionally, it will promote collaboration and inform the IYS Data Scientists whether new vocabulary terms will have to be developed in an appropriate controlled vocabulary repository.

2.2.2 Metadata Record Creation and Review

While there will be one Data Management Plan per Research Area, IYS Data Scientists will generate a metadata record for each data set based on information contained in the Data Management Plan. A 'data set' will be defined by each distinct sampling platform and the associated research vessel and research team to capture the heterogeneity in sampling protocols and data attribution. Using this approach, the metadata provided is consistent, accurate, and compliant with OBIS and GOOS standards. The IYS Metadata Catalogue will be public. However, metadata records will remain private until verified by the IYS Expedition Scientists. Metadata records will be created as soon as practical, ideally in advance of data collection and publishing. Data access will be added to metadata records after data processing and publishing is completed. A timeline for this process must be submitted and agreed to in writing when submitting Data Management Plans.

2.2.3 Data Processing

Data collection protocols and standards will vary between vessels based on their historical standard operating procedure or reporting and data storage requirements. It is, therefore, not possible to define one workflow for every vessel that covers procedures spanning data collection to standardization. Each vessel will have their own 'Raw Data' format (see definition in methods section), and a varying number of steps required to produce a 'Processed Dataset' to share with the IYS Data Scientists. There is no attempt to address or standardize the various steps that will be unique to each vessel related to collecting, entering, quality checking, or storing 'Raw Data'. Documentation on those procedures is expected, however, when providing the Processed Data.

Once the 2022 salmon ocean ecology data are collected and processed, IYS Expedition Scientists will be asked to provide their Processed Data that, where applicable, adheres to international standards and contains controlled vocabulary. The IYS Data Scientists are developing a Practical Data Standardization Guide and Data Dictionary to be distributed prior to the 2022 Expedition to define data standards to ease integration with global data repositories. After this guide is distributed, it is recommended that a workshop / training session be conducted by the IYS Data Scientists for those responsible for processing data and submission. Once data are processed, the IYS Data Scientists should be given access to this Processed Data in one of several ways. In order of preference:

1. Access is provided to the National or Institutional Data Storage platform (eg. FTP server, ERDDAP server, SQL database, Google Drive, GitHub) that stores Processed Data so that Hakai Data Scientists can query this database programmatically to ensure any updates to data are easily captured and version control is adequate;
2. Processed Data (with a version identified) can be emailed directly to iys.data@hakai.org and subsequently hosted on the private IYS GitHub repository[2].

2.2.4 Data Standardization and Publishing

For biological data such as species occurrences, gut contents, and fatty acids IYS Data Scientists will transform dataset structure to the OBIS ENV-DATA format, which will create a relational data model with a nested structure of events, species occurrences and measurements or facts related to events or occurrences. Using this method, individual datasets can be easily related to each other in a coherent data model that represents how the data were collected and how they are related to each other.

Data falling under the BioGeoChemical or Physical EOV domain of GOOS, will be uploaded to ERDDAP because this is the most widely adopted data access platform for this type of data. Hosting data on an ERDDAP server will make the data available through a number of different methods including an html link from the IYS Metadata Catalogue directly to the dataset hosted on ERDDAP, through other ERDDAP data servers, as well as through an application programming interface (API) that can be used to programmatically access data. This can be done using tools such as R packages, Python packages, or stand-alone programs can be built in Java, for example, which can access data through an API. Data hosted on ERDDAP will use the same event IDs used in the biological data sets on OBIS so that data remain easily related but will not strictly adhere to the OBIS-ENV-DATA format.

2.2.5 Data Access and Sharing

The CKAN IYS Metadata Catalogue will serve as this single point of entry data portal to all the standardized IYS data and each metadata record will link to a data set either in OBIS or ERDDAP (see Figure 2). These platforms do not require a user to log in or request access to data, making the data openly accessible to the broader scientific community. Metadata in the IYS Metadata Catalogue can be harvested by the GOOS-affiliated platforms, such as the Canadian Integrated Ocean Observing System (CIOOS) and IOOS (the American equivalent to CIOOS). As GOOS-affiliated platforms use consistent metadata formats and protocols for data exchange, the salmon ocean ecology data that feed into these platforms is findable, accessible, and in some cases interoperable with the multitude of other datasets in this framework. This makes the data easier to discover and, more importantly, scientifically comparable with other data that have adopted these international standards. The Federated Research Data Repository (FRDR) further harvests metadata from these platforms and acts as a back-up.

[2] https://github.com/orgs/international-year-of-the-salmon/

In addition to the Standardized Data, the IYS Metadata Catalogue will provide access to the Processed Data submitted by data providers and the required supplementary materials including a README file giving an overview of the data set and any methods description documents by linking to either the National or Institution Data Storage platform, or optionally to a private IYS GitHub repository. This private repository is only accessible to the responsible party (scientists, organization) and the IYS Data Scientists (data custodian) during curation and standardization, ensuring security of the data while simultaneously ensuring version control. This platform could be utilized for sharing data internally with other IYS Expedition Scientists before the data are made public. Storing the Processed Data and supplementary materials in this GitHub repository has multiple advantages compared to using other ad-hoc data storage platforms:

- Data can be stored in a private repository, accessible only to the scientist, the data custodian, and other collaborators as needed. Once the repository is made public, or the data moved to a publicly accessible repository, the data adhere to the FAIR data principles.
- This data repository provides version control for the data, making it very easy to see what version is the latest and what the changes are between versions. This ensures that the data standardization is done to the latest version and updates and corrections to the original Processed Data can easily be integrated into existing analyses.
- The standardization R and Python scripts used to quality check and standardize the Processed Data are also stored in the repository, allowing scientists and users to audit data transformations.
- GitHub offers a platform for communication, and questions related to the (meta)data can be asked directly to scientists if tagged in ‘Issues’.

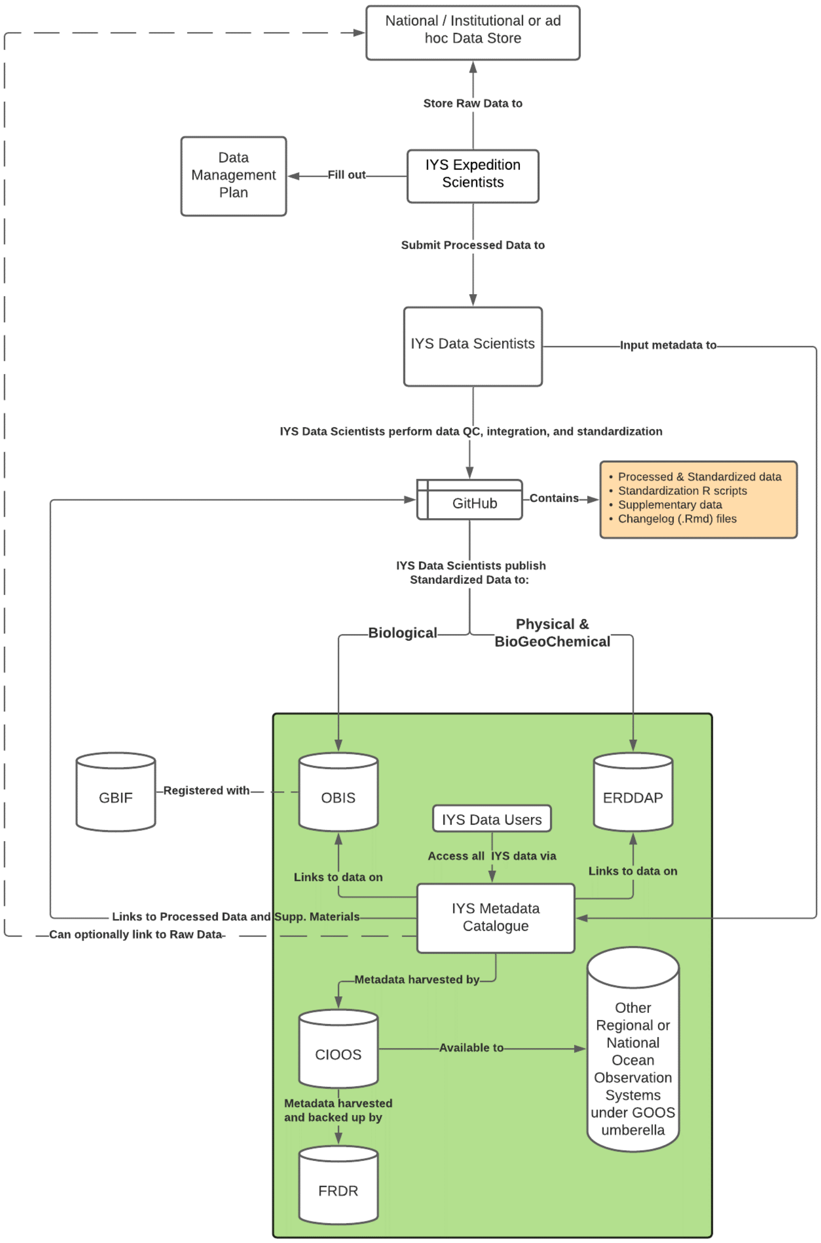
National / Institutional or ad hoc Data Store
Store Raw Data to
Data Management Plan
Fill out
IYS Expedition Scientists
Submit Processed Data to
IYS Data Scientists
Input metadata to
IYS Data Scientists perform data QC, integration, and standardization
GitHub
Contains
• Processed & Standardized data
• Standardization R scripts
• Supplementary data
• Changelog (.Rmd) files
IYS Data Scientists publish Standardized Data to:
Biological
Physical & BioGeoChemical
GBIF
Registered with
OBIS
ERDDAP
IYS Data Users
Access all IYS data via
Links to data on
Links to data on
Links to Processed Data and Supp. Materials
Can optionally link to Raw Data
IYS Metadata Catalogue
Metadata harvested by
CIOOS
Available to
Other Regional or National Ocean Observation Systems under GOOS umberella
Metadata harvested and backed up by
FRDR

Figure 2: Conceptual flow chart describing the different data platforms and processes throughout the proposed data mobilization process. Discipline-specific metadata associated with the IYS data is stored in the IYS Metadata Catalogue, which includes links to both the Processed and Standardized Data.

3. Proposed Data Mobilization Communication Strategy

A clear and well-defined communication strategy related to data mobilization is essential to having an organized workflow throughout the IYS project. Initially, communication regarding the data mobilization and communication process could be done between the IYS Data Scientists, the IYS Secretariat, the IYS Cruise Planning Team and the NPAFC Study Group, created ad hoc by the Committee on Scientific Research and Statistics (CSRS) (NPAFC, Doc. 1926).

Standardization of the 2019 and the 2020 data collected will help identify key areas of attention, roadblocks, and help structure the Data Management Plan and standardization workflow for the 2022 multi-vessel Pan-Pacific Winter High Seas Expedition. Through our initial communication sessions, uncertainties can be addressed and feedback incorporated regarding the data mobilization and standardization approach.

To increase participation and understanding of the data mobilization cyber-infrastructure and standardization process, it is recommended that this process, once finalized, is discussed and presented to the IYS Expedition Scientists involved in each Research Area. Using a case study and demonstrating the process and application of data standardization and publishing could increase collaboration efforts and engagement of scientists. Furthermore, it is recommended that a workshop or training session be conducted by the IYS Data Scientists for those responsible for processing data and submission, covering the details in the forthcoming Practical Data Standardization Guide. These presentations can help set the stage for future communications between the IYS Data Scientists and IYS Expedition Scientists involved, should any questions regarding the data arise post-expedition. Active feedback and scientist participation and engagement will help construct and further develop the Data Mobilization and Communication plan and Data Management Plans for future expeditions.

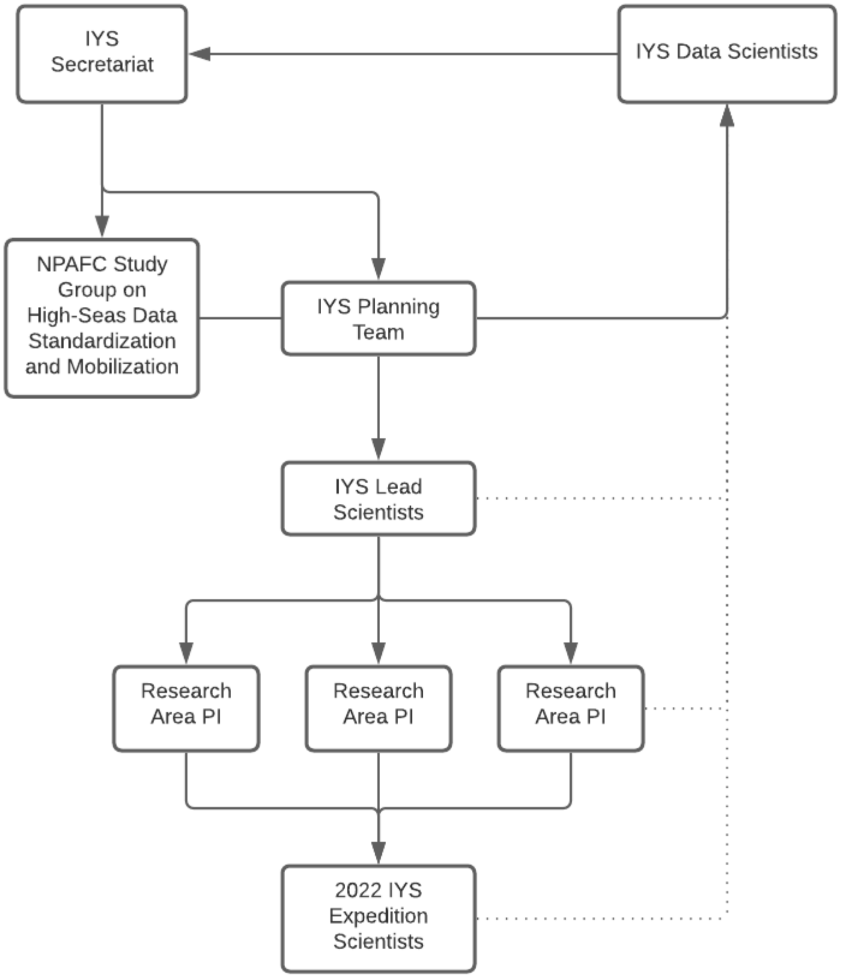

Figure 3: Proposed communication flowchart. Initial communication on finalizing the data mobilization and standardization process will be done between the IYS Data Scientists, IYS Secretariat, Study Group, and the IYS Planning Team (solid lines). Once finalized, it will be communicated down the chain, giving IYS National Lead Scientists, PIs and 2022 Expedition Scientists opportunity to give feedback and ask discipline-specific questions regarding the process to the IYS Data Scientists (dashed line).

4. Conclusion

This Data Mobilization and Communication plan describes the data standardization approach of the salmon ocean ecology data collected during the IYS research cruises. Integrating and reformatting the data to an international standard increases the longevity and interoperability of the data and ensures that the data adhere to the FAIR data principles.

Furthermore, a brief overview is provided on the standardization, mobilization and integration workflow. A communication scheme is proposed on how best to distill the data mobilization and standardization process through the various layers of the IYS project.

5. Important links and resources

- OBIS: https://obis.org/ – https://obis.org/manual/
- NERC Vocabulary: https://www.bodc.ac.uk/resources/products/web_services/vocab/
- GitHub IYS Hakai: https://github.com/HakaiInstitute/iys-oos
- GOOS Framework for Ocean Observing: http://www.oceanobs09.net/foo/FOO_Report.pdf

6. References

- De Pooter, D., Appeltans, W., Bailly, N., Bristol, S., Deneudt, K., Eliezer, M., ... & Lipizer, M. (2017). Toward a new data standard for combined marine biological and environmental datasets-expanding OBIS beyond species occurrences. Biodiversity Data Journal, (5).
- Stewart, A., Deyoung, B., Smit, M., Donaldson, K., Reedman, A., Bastien, A., ... & Plourde, A. (2019). The development of a Canadian integrated ocean observing system (CIOOS). Frontiers in Marine Science, 6, 431.
- Tanhua, Toste, Sylvie Pouliquen, Jessica Hausman, Kevin M. O'Brien, Pip Bricher, Taco De Bruin, Justin James Henry Buck et al. "Ocean FAIR data services." Frontiers in Marine Science 6 (2019): 440.