

# IYS Data Policy

Oct 13th, 2021

The 2022 International Year of the Salmon (IYS) Pan-Pacific Winter High Seas Expedition is a collaborative, international project to address the scientific concerns around salmon marine survival in the Pacific Ocean. The project's success, and its ultimate impact on science and society, relies upon professional coordination and data sharing across the project and the broader scientific community. A transparent Data Policy is essential to achieve the IYS science objectives, to facilitate collaboration, and to enable broad use and deepen the impact of the IYS data.

## Executive Summary

This Data Policy regulates data management, access and release, as well as authorship and acknowledgment. Signing this Data Policy is a **prerequisite** for participation in the 2022 International Year of the Salmon (IYS) multi-vessel pan-Pacific Winter High Seas Expedition (hereafter “**the Expedition**”, or “**2022 IYS Expedition**”).

## 1. Objective

The purpose of this Data Policy is to codify the goals and principles of the IYS research data lifecycle from production, documentation, sharing, usage and re-usage. This ensures that procedures for data gathering, archiving and publication, as well as metadata and quality management are commonly implemented. By participating in the IYS project, all IYS Participants agree to comply with this Data Policy. By doing so, Participants ensure that the IYS is a successful and resource-effective research project that also supports data accessibility, interoperability and re-usage following the FAIR data principles.

This Data Policy aims to:

1. Provide clear expectations to expedition scientists and organizers regarding data access and release, as well as authorship and acknowledgment.
2. Ensure the fair and equitable use of IYS data and uphold the rights of individual scientists and institutions.
3. Encourage the rapid publication and dissemination of scientific data, results and knowledge, to support the involvement of a broad user community.

## 2. Definitions

- **IYS data:** Salmon ocean ecology data collected, measured, recorded, created or derived throughout the 2022 IYS Expedition. This includes biological and biogeochemical data,

derived data from analyzed laboratory sample material, sample and event metadata, satellite data products and data collected from Argo floats or autonomous glider(s).

- **Data provider/PI:** All data streams must have a responsible party. The data provider/PI is responsible for the data quality and delivery. The data provider is defined as the PI or institution that owns and/or operates an instrument, creates and/or analyzes samples, produces a model output, or otherwise produces a data set.
- **IYS Participants:** Participants of the Expedition whose scientific activities are officially endorsed by the IYS Steering Committee. Participants are bound to the IYS Data Policy and will have access to the processed IYS data before public release.
- **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. It is data directly produced by sensors, devices, or manual observation, prior to additional processing, calibration and quality assessment/control (never modified).
- **Processed data:** Data that have had basic event and metadata quality assurance and result from data providers fitting their Raw data into the IYS provided data template spreadsheets.
- **Derived data:** Products based on raw data that may involve derivation of additional parameters or delayed-mode quality control using external data or post-use sensor calibration; model data or a combination with any external data, eg. by data assimilation, visualization, classification, or clustering.
- **Synthesis data:** Data from the IYS High Seas Expeditions combined with already published, historic or model data.
- **Data Packages:** A .zip file containing the processed data (.xlsx or .csv format), a data dictionary with definitions pertaining to the (meta)data (measurement) attributes, sampling protocols, calibration files and documentation of eg. processing steps.

### 3. Metadata Standards

Metadata shall make data findable and provide additional contextual information about measurement details, methods, relevance, lineage, quality, usage and access restrictions of the data. The metadata should be agreed on and listed in the Data Management Plan (DMP).

All variables and parameters (measurement attributes) must be documented with an attribute name and attribute definition that provides a human-readable context for the measurement. For numeric data, attributes must include the units of measurement using SI unit definitions. For numerical data without a unit (e.g., percent, count x per count, etc.) the unit should be noted as "dimensionless". For non-numeric, categorical data, coded values must be defined in a code/definition list, or be defined by an external, controlled vocabulary term. The [NERC Vocabulary Server \(NVS\)](#) web service provides access to controlled vocabularies via an international, actively-contributing research community. Any deviations from this recommendation, or from using SI unit definitions, must be documented in the DMP.

#### 4. Data Ingest, Transfer, Storage and Archiving

Raw and processed salmon ocean ecology data may be stored on national or institutional data storage platforms, as defined in the DMP. These data storage platforms should be able to archive data over the long-term and provide unique and stable identifiers for the datasets. The IYS Data Scientists from the Hakai Institute must be provided access to these platforms to harvest processed data. Processed data in these data storage platforms needs to be clearly versioned. Other IYS Participants might be provided access by the data provider upon request. If data cannot be stored on national or institutional databases due to institutional data policies or access cannot be provided to the IYS Participants due to data policies or commitments to other stakeholders, a **data package** should be sent to the IYS Data Scientists. The [2022 International Year of the Salmon GitHub repository](#) will be used by the IYS Data Scientists for storage of the data packages and/or harvested, processed data from the national or institutional storage platform. Only IYS Expedition Scientists, Country Lead Scientists, and the IYS Data Scientists with authentication/authorization will have access to the data in this repository prior to public release. Authorization to this secure GitHub can be requested from the IYS Data Scientists.

In any case, each dataset must have a clearly identified primary archive as stated in the DMP. Any deviations from the rules and guidelines stated here need to be agreed on between the data provider and the IYS Steering Committee, and communicated with the IYS Data Scientists.

The [IYS Metadata Catalogue](#) will provide a centralized point of access and reliable, long-term storage for the metadata records for each individual data package produced. The metadata will be derived from the DMPs. Metadata records will be stored during and beyond the duration of the IYS project. Metadata is ingested to the IYS Metadata Catalogue ahead of the Expedition by the IYS Data Scientists, and will be publicly visible. Once IYS data are processed, standardized and mobilized to the relevant global data repositories, the IYS Data Scientists will update the metadata records in the IYS-OOS to include a link to the processed and standardized data. Throughout this process the metadata record can be revised and versioned.

The responsibility for the accuracy of the metadata record lies with the data provider.

#### 5. Data Provision and Sharing among the IYS Participants

Early access by the IYS Participants to the data is crucial for successful collaboration. Hence, all processed data must be made available to all IYS Participants as fast as possible. Data and metadata included in the Cruise report following the Expedition should become publicly available at the same time as the report is published. Processed data shared with the IYS Data Scientists will be stored on the 2022 IYS GitHub repository, and become freely available to all IYS Participants. Through automation of data harvesting from the national or institutional data storage platform, the IYS Data Scientists will ensure that the latest version of processed data is stored on the repository. If processed data is sent to the IYS Data Scientists in a **data package**, it is the responsibility of the data provider to ensure that the latest version is provided.

For raw data archived on a national or institutional data storage platform, IYS Participants are encouraged to connect directly with the **data provider** to request access or ask questions. Raw data will not be stored on the 2022 IYS GitHub repository. The **data provider** is responsible for ensuring raw data longevity.

Table. 1 Description of IYS data categories and their schedule of availability to IYS participants and public release.

Data Category	Example datasets	Availability	
		To Participants*	To Public
At sea observations	Fish trawl, bridge log, wildlife observations, physical oceanography, gut contents, CTD, YSI	Immediate after basic QA/QC	6 months
Sensor data	Argo float / glider	Immediate after basic QA/QC	6 months
Post expedition laboratory sample analysis	Stock ID, pathogen	As soon as available	2 years
Satellite	Chlorophyll a, radiance,	As soon as available	2 years
Metadata	All datasets	Before cruise	Immediately after cruise and review from data provider

\* Via private IYS GitHub data repository

The following deadlines mark the latest points in time for transferring processed data to the IYS Data Scientists, or providing the IYS Participants access to processed data or data packages stored on national or institutional databases:

- Sensor data: This includes e.g., CTD data. Sensor data should be made accessible to IYS Participants no later than **date xxx**.
- Bridge Log data: Information related to sampling station coordinates, sampling time, environmental conditions and other metadata should be made accessible to the IYS Participants no later than **date xxx**.
- Laboratory sample analysis data: This includes eg., nutrient data, genetic samples and derived data. These data must be made accessible to the IYS Participants no later than **date xxx**.
- Trawl catch data: These data must be made accessible to the IYS Participants no later than **date xxx**.

- Satellite data: This includes eg. satellite chlorophyll-a data. Must be made accessible to the IYS Participants no later than **date xxx**.
- Historic data: This includes eg. historic fish trawl data. Must be made accessible to the IYS Participants no later than **date xxx**.

Exceptions to the processed data delivery deadlines can be made but should be discussed and agreed on between the **IYS Expedition Scientist(s) and the Steering Committee ahead of the Expedition**.

For using data collected during the 2022 IYS Expedition before data are published publicly, the **data provider** must be informed and offered collaboration on the scientific analysis and must be offered co-authorship based on the principles described in section “Authorship and Acknowledgement” below. The **data provider** may object to the usage of data in a publication if that publication conflicts with his or her own publication strategy. Any such objection must be discussed and agreed upon in writing with the **IYS Steering Committee**. The **data provider** may not object to the usage of data beyond the public release date.

#### 4. Public Release of Data

Good progress of a highly collaborative and interdisciplinary project like IYS requires open availability of data to a wide user audience as early as possible. At the same time, it is important to acknowledge the substantial work that goes into collecting, quality controlling, formatting, documenting, and releasing scientific data. IYS policies pertaining to data use and acknowledgment aim to balance these two principles. Data access and usage policies evolve in time according to a staged process outlined here, and in all cases the most data-restrictive approach is described while an accelerated publication of data is acceptable.

Processed IYS data will become freely and publicly available on the 2022 IYS GitHub repository and on global data repositories on **March 31, 2023**. On this date, the GitHub repository will either be made public, or folders will be transferred to a public repository. Additionally, the metadata records in the IYS Metadata Catalogue will be updated to include a link to the publicly visible processed data. From this date on, there will be no restrictions on data usage aside from the [Creative Commons Attribution licensing](#), but data users are strongly encouraged to communicate with **data providers** during early stages of all scientific analyses to ensure accurate usage and interpretation of data.

Prior to publication and mobilization to the global data repositories, the processed data will be (further) standardized if required, to ensure interoperability with the multitude of other datasets available in those repositories.

#### 5. Authorship and Acknowledgment

Generally, **co-authorship** on publications and other public documentation must be offered to those that have **made a substantial contribution** following the [Code of Good Scientific](#)

**Practice.** An inclusive co-authorship approach is encouraged. Co-authorship on publications and other public documentation must generally be offered to those that a) have made a substantial contribution to the creative process, that is, to the conception and design of the study, or to the analysis and interpretation of the data (i.e., including the **data provider**); b) have contributed to the preparation of the communications, reports, or publications that have arisen; c) be able to present in detail his or her contribution to the project and to discuss the main aspects of the overall research. All authors should confirm in writing their agreement with the final version of original manuscripts submitted for publication. Lead authors have the ultimate decision authority and responsibility to identify and appropriately engage co-authors. Additional standard guidelines for deciding on co-authorship on publications can be found via numerous on-line resources, such as <http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html> or <https://www.dfg.de/sites/flipbook/gwp/files/assets/basichtml/page85.html>.

Contributors to the work that do not warrant co-authorship should be identified by name in the acknowledgments. Authorship conflicts may be resolved by the **IYS Steering Committee**.

IYS data must be acknowledged or referenced in publications and other public documentation, specifically including relevant digital object identifiers (DOIs), data providers (if not co-authors), and funding agencies.

## **6. Data Publication**

Clear, consistent documentation and standardization of IYS data will help to support a strong and lasting IYS data legacy, promote the broad and appropriate use of IYS data including the citation of data, and ensure proper acknowledgment of data creators. The publication and mobilization of IYS data to global data repositories will be supported by the IYS Data Scientists. To this end, IYS Expedition Scientists will be provided standard data spreadsheet templates ahead of the 2022 IYS Expedition to facilitate project data integration using consistent data table formats, columns and structures. Not only will these templates ensure consistent data recording among vessels, but it will also permit the IYS Data Scientists to create scripted data transformations to international standards required by global data repositories.

The ultimate goals for data publication are to ensure IYS data interoperability at both the project and global scales so that IYS data can be integrated with other global efforts. To assist in data tracking, awarding of credit and data provenance, IYS Data Scientists will create a digital object identifier (DOI) for each metadata record.

Authorship on data publications should follow similar policies to authorship on scientific publications and must include those participants that have made substantial contributions to collecting the data, processing the data, and documenting the data. Each data publication needs a contact person (typically the data provider) and PI who is familiar with and responsible for the scientific evaluation.

IYS data may serve as a basis for synthesis data products, i.e., data from IYS combined with already published, historic or model data. Synthesis data should be published in the same manner as IYS data. PIs working on synthesis data and related publications are encouraged to ensure that data from other sources that are a part of synthesis data are published. At minimum, metadata records for historic or synthesis data should include links to various other sources of data used.

## 9. Amendments

**Dispute resolution:** Disputes on the Data Policy should be solved primarily by the involved individuals or [ ].

### Signature

<b>Name</b>	
<b>Institution</b>	
<b>Email</b>	

Hereby I declare that I fully consent to the IYS Data Policy and become a registered IYS Participant.

-----

Date, Signature