International Year of the Salmon Data Mobilization Recommendations

March 31, 2020

Brett Johnson

Hakai Institute

1713 Hyacinthe Bay Road, Heriot Bay, BC, Canada

Executive Summary

An agreement was signed between the Hakai Institute and the North Pacific Anadromous Fish Commission on February 3rd, 2020 for Hakai to scope and review the requirements of data management for high seas salmon ecology and oceanographic observations collected by the International Year of the Salmon collaborative project. We outline a number of recommended actions and considerations for building and delivering digital infrastructure systems for inetgrating a keystone data ecosystem–salmon ecology. Foremost we recommend timely, inclusive, and equitable data access. The Global Ocean Observation System is a project of the United Nations Educational, Scientific and Cultural Organization and unifies networks of scientists around the world. Adopting the international standards and principles outlined broadly by UNESCO will ensure a multilateral approach inline with the UN Declaration on the Rights of Indigenous Peoples to the standardization, integration, and equitable distribution of salmon ocean ecology data in British Columbia and beyond.

Data collected by the 2019 research cruise is centrally accessible in an International Year of the Salmon Ocean Observation System (IYS-OOS) catalogue at https://iys.hakai.org. We provide a complete catalogue of the data sets produced by the IYS in 2019 with ISO 19115 compliant metadata records, making the knowledge that these data exist publicly discoverbale-a fundamental first step.

We reccommend four components of a Data Management and Communications Model: 1) Data catalogue records compliant with ISO 19115 (http://iys.hakai.org)-- COMPLETE; 2) Open-Access licensing and Open Data Access Protocols; 3) Controlled Vocabularies that define the variables, methods, units, platforms and measurement types used in salmon ocean ecology adhering to 'Ocean Best Practices' maintained by the Global Ocean Observation System; 4) A dedicated TRUSTed digital repository for hosting data and data-analysis tool-development code securely in perpetuity.

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. Members nations of the NPAFC are collaborating on a multi-vessel Oceanographic Expedition planned for March

20201 covering from California North and West to Kamchatka and as south as South Korea, including the Sea of Ohkotsk and parts of the Berring Sea planned for March 2021. Transdisciplinary researchers spanning Physical, BioGeoChemical, and Biodiversity/Ecosystem domains from at least a dozen institutions and agencies will generate a complex set of data. Success will be measured by timely and equitable access to data and knowledge generated by the International Year of the Salmon. The NPAFC and the Hakai Institute with support from the British Columbia Salmon Restoration and Innovation Fund and the Tula Foundation are conducting a review of current practices and new approaches to mobilizing salmon ocean ecology data, specifically for the data collected during the five-vessel survey planned for 2021.

Methods

Overview

For every data element, collection method, platform, and variable produced by the IYS High Seas Expeditions the following tasks need to be completed:

- Determine whether the data element is already defined within GOOS framework. Such elements will be processed first because the requirements are well-defined. For data elements that do not naturally belong in IYS-OOS, determine whether there is a recognized and compatible repository where they belong and can be federated or linked to the IYS-OOS. Example alternative data repositories: Biodiversity of Life Online Database (BOLD), DataONE, Driad, Federal Open Data, BC Gov. Data etc.) TODO: improve previos egs.
- Publish. For all data elements, generate appropriate and valid metadata records to make the existence of the data public knowledge and insert the records into the metadata catalogue on the IYS Data Portal, so that they are discoverable by IYS users.
- Process. Work closely with the data provider to bring fully validated and standardized copies of data elements into the appropriate repositories.
- Communicate. Representatives from each scientific discipline involved with the IYS should connect through a working group that disseminates and advocates for best-practices.

Planning

For the 2021 cruise to be successful, the establishment of a data standards study team made up of relevant representation of stakeholders needs to be established as soon as possible to begin to prioritize every data element, method, platform, and variable they plan to collect. The NPAFC is a natural platform for this study group, so is the adoption of data standards put forward by the Global Ocean Observation System (GOOS). GOOS is a program that is coordinated by the Intergovernmental Oceanographic Commission of UNESCO (United Nations Educational, Scientific and Cultural Organization). GOOS is governed by a multinational Steering Committee, three scientific domain Expert Panels, and many Observation Coordination Groups of people and organizations worldwide. GOOS is partnered with expert agencies

in biological data—namely the Ocean Biogeographic Information System (OBIS), Biodiversity of Life Online Database (BOLD) and the Marine Biodiversity Observation Network (MBON), Intentational Ocean Observation System, and the Canadian Integrated Ocean Observation System. We recommend using established international standards connected to GOOS where available, and extending or developing standards where needed. The product of this strategic alignment and development we can call the International Year of the Salmon Ocean Observation System (IYS-OOS) for now.

Data Exchange Protocols

Communications Plan

Data Strategy

There needs to be a long-term strategy to effectively engage important stakeholders and deepen the impact of data mobilization. Community development is critical for the IYS-OOS to be successful. Open Science. Repeatability. Reproducibility. FAIR Data. But hold on, what is copyright? Public work? Who is the public? Equality and openness; major re-balance of power required. UNDRIP. Deepening Engagement and Impact; 2) System Integration and Delivery; and 3) Building for the Future.

Steering Committee

An executive-level steering committee will provide technical and strategic advice on the project while Hakai and NPAFC will retain administrative oversight of the project. Steering Committee participants will include:

- Eric Peterson/Brett Johnson – Hakai Institute
- Mark Saunders/Stephanie Taylor– NPAFC
- Dick Beamish and Brian Riddell – 2019 and 2020 Expeditions/Pacific Salmon Foundation
- Bruce Patten – DFO Pacific Biological Station and OBIS Canada Node Manager
- Erin Satterthwaite – GOOS BioEco Expert Panel Member
- Brian Hunt – University of British Columbia Institute of Oceans and Fisheries, Expedition Scientist
- Expedition Chief Scientists

Development Groups

The development of the data model will rely heavily on input from expedition scientists, project affiliated Scientists, Professors, and others. Within each scientific domain planned for the 2021 cruise, there should be representation in an ad hoc working group tasked with determining international data standards that apply to their domain of expertise

Hakai Staff and affiliates to consult as needed:

- Eric Peterson. Strategic direction.
- Ray Brunsting. Hakai Chief Technology Officer.
- Matt Foster. Hakai Chief Data Architect.
- Jennifer Jackson. Physical oceanographer for Hakai.
- Brian Hunt. A faculty member at UBC and the head of Hakai salmon program

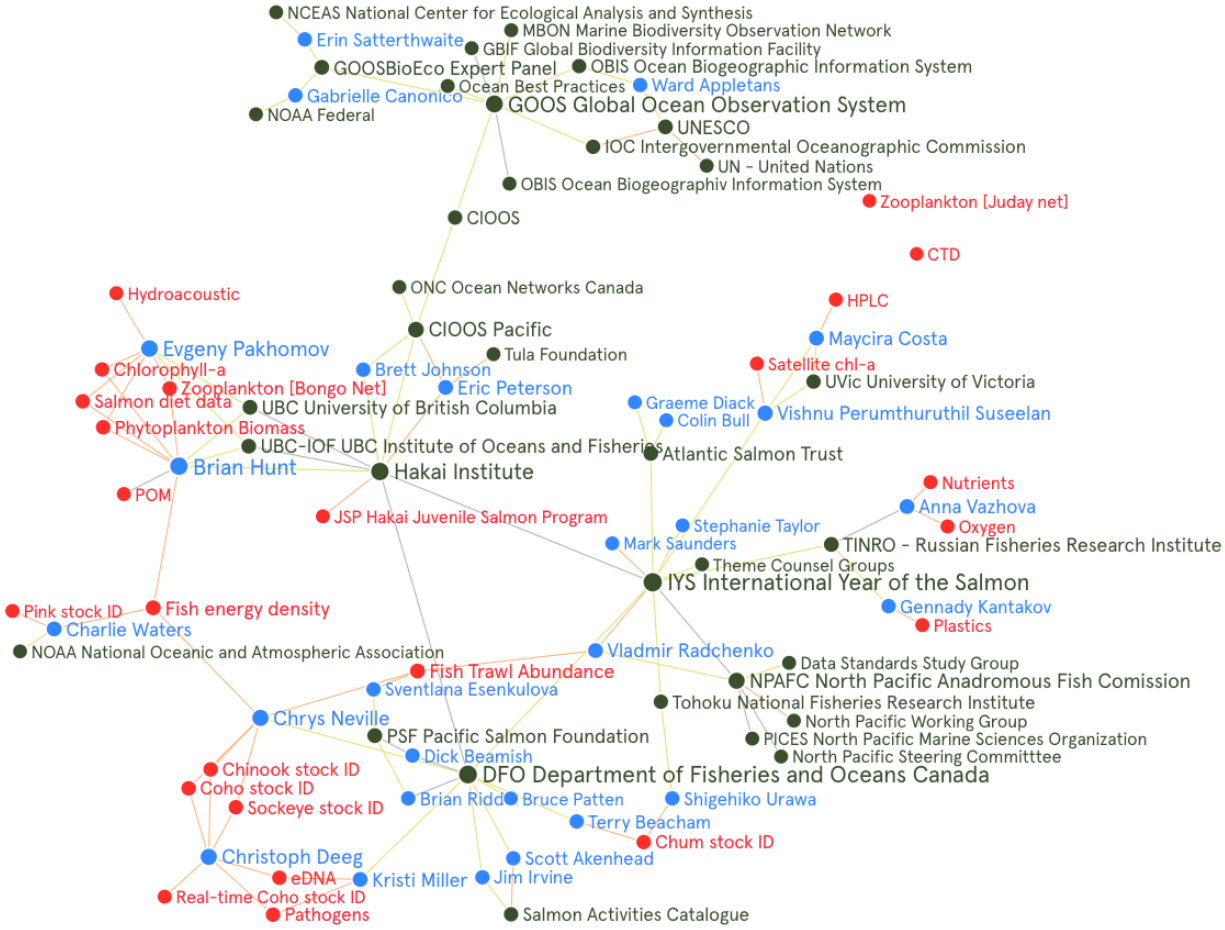

Figure 1: Graph network of the people, organizations, and data related to the International Year of the Salmon 2019 Expedition

Timeline

Here is the proposed road map and timeline which will be refined before finalization for May 30, 2020.

TODO: Update with Release report, backlog report, and sprint backlog.

- Produce a comprehensive Data Strategy May 30, 2020
- Begin Building Data Management and Communication Model June 1, 2020

- Establish 2021 cruise Data Standards Group ASAP
- Bring 2019 Cruise Data into the Correct Repository(s) August 2020
- Bring 2020 Cruise Data into the Correct Repository(s) September 2020
- Identify leaders for 2021 Cruise Data Management ASAP 2020
- Integrate 2019, 2020, and 2021 cruise data TBD 2021
- Extend GOOS framework into Salmon Ocean Ecology and Integrate with Salmon Activities Catalogue TBD 2021

Review

Backlog

Milestones

Sprints

Time Sensitive Challenges

Some methods used in the field need better documentation in the metadata associated with each dataset produced on the expedition. This includes equipment and platform descriptions, calibration files, instructions for how data were summarized or aggregated, and any caveats for data interpretation etc… This will help in ensuring the scientific integrity of the consolidated data sets. Dataset 'Quality' levels will eventually be assigned based on completeness of metadata collection and integrity of provenance. Development of these guidelines among research domains is pressing.

Trawl catch data has not yet been received. My current understanding indicates that the dataset structure is under development, and that expedition Scientists are working on this. Please understand that access to raw outputs and the detailed processing steps that occur to transform data are needed for complete data provenance. Keeping track of changes to raw data will help us ensure reproducibility, which is becoming commonly required in life sciences journals despite it being a high bar to aim for. Using a change log that you maintain manually ensures that reproducibility can be achieved when data cleaning is performed ad hoc such as removing outliers during quality control procedures, etc. Otherwise, we recommend moving to using scripted data transformations and distributed version control to help scientists collaborate on common data and analyses.

There are also a number of questions we have related to specific data sets that we keep track of here for data providers to view and respond to. We recommend the ad hoc development of data standards group ASAP. A core group comprised of representatives of each research domain (physical oceanography, salmon feeding ecology etc.) and representation from as many member nations, and stakeholders as feasible.

Produce a Data Strategy Document that will:

- Identify project milestones for 2020, 2021, and 2022 and requisite steps to success

- Describe the foundation for a new best-practice approach to provide open and FAIR access to salmon, oceanographic, and climatic data integrated from numerous sources.
- Propose a survey of key data providers to deepen engagement.

Begin Building Data Management & Communication Framework

Start building a data management and communication framework that initially comprises:

- A web-based IYS Data Portal. We propose to follow the lead of DFO/CIOOS and implement a cloud-based solution with Amazon Web Services, utilizing the Montreal data centre to host a data catalogue/portal.
- A metadata catalogue (we propose to follow DFO/CIOOS and employ CKAN), which allows for data discovery across the federated system.
- A GOOS-compatible repository for physical and biogeochemical data acquired on the expeditions.
- Solutions for other data types: either as natural extensions of the GOOS model, as contemplated under the GOOS Bio Eco, or in other repositories, as mandated by our modelling decisions.
- Concentrate first on the data types acquired during the 2019 cruise.

Links and Resources

- IYS-OOS GitHub Repository
- Please comment on this document issues in github
- Temporary 'AirTable' IYS-OOS Database
- Template CKAN Catalogue
- Global Ocean Observation System
- OBIS ENV-DATA Darwin Core Archive Data Structure
- Good enough practices in scientific computing