International Year of the Salmon Data Mobilisation Strategic Recommendations

Hakai Institute

Brett Johnson, Tim van der Stap

1713 Hyacinthe Bay Road, Heriot Bay, BC, Canada

Last Updated: 2020-09-29 13:27:44

Executive Summary

An agreement was signed between the Hakai Institute and the North Pacific Anadromous Fish Commission on February 3rd, 2020 for the Hakai Institute to scope and review the requirements of data management and data mobilisation for salmon ecology and oceanographic observations collected by the International Year of the Salmon High Seas Research Expeditions in the North Pacific Ocean. Here, we recommended a number of actions and considerations for building and delivering cyber infrastructure systems to integrate a keystone data ecosystem – salmon ocean ecology data. Foremost we recommend timely, inclusive, and equitable data access under FAIR terms (Findable, Accessible, Interoperable, and Reusable). To that end, we recommend adopting the Global Ocean Observing System (GOOS) put forth the United Nation's Educational, Scientific and Cultural Organization. Adopting the best practices for ocean data management outlined broadly by UNESCO will ensure a multilateral approach to the integration, standardization, and distribution of salmon ocean ecology data.

We recommend four components of a Data Mobilisation Model: 1) Data catalogue records compliant with ISO 19115 (http://iys.hakai.org); 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; 4) A federated, dedicated, and trustworthy system of data repositories for hosting salmon ocean ecology data and data analysis tools securely in perpetuity.

Data sets collected by the 2019 High Seas Expedition are centrally Findable in an International Year of the Salmon Ocean Observing System (IYS-OOS) catalogue at https://iys.hakai.org. We've built and delivered the data catalogue infrastructure, including ISO 19115 geospatial-metadata-standard compliant data-catalogue-records, to facilitate mobilising privately-funded data produced in 2019 and 2020 and for data produced in 2021 and beyond. We continue to facilitate data mobilisation for historical data sets, including 2019 and 2020, but the onus has now shifted to project-associated Scientists, Principal Investigators, and their funders to determine how to best fit their historical data into the International Year of the Salmon Ocean Observing System. A major focus for ongoing data mobilisation work is the automation of data handling procedures and the development of human-friendly data-management work flows, best-practices, and communication protocols.

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. Member nations of the NPAFC are collaborating on a multi-vessel Oceanographic Expedition planned for March 2022 covering from California North and West to Kamchatka, Russia and as far south as South Korea, including the Sea of Okhotsk and parts of the Bering Sea. Trans-disciplinary research spanning Physical, Biogeochemical, and Biology/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 these expeditions. 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 mobilising salmon ocean ecology data, specifically for the data collected during the multi-vessel surveys planned for 2021 and/or 2022.

Methods

Data Mobilsation Model

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

1) Define. Determine whether the data element is already defined within the GOOS Essential Ocean Variable framework. For data elements that do not naturally belong in the GOOS EOV framework, determine how third-party definitions of data elements might be synonymous with GOOS EOVs. In cases where collected variables are not synonymous with GOOS compatible terms, determine whether there is a recognized and compatible repository where data belong and can be federated or linked to the IYS-OOS. Example alternative data repositories: Barcode of Life Database (BOLD), the Ocean Biodiversity Information System, Global Biodiversity Information Facility, DataONE, Federal Open Data, BC Government public data etc.)
2) 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 'Findable' and 'Accessible' by IYS data users.
3) Transform and Integrate. Work with data providers to transform data copies, subsets, or aggregates of data elements so they can be integrated into appropriate repositories identified by domain expert scientists, making the daya 'Interoperable'
4) 'Communicate. Implement flexible Data Mobilisation and Communication Plans that can be adapted to the needs of specific science domains to facilitate data Re-use.

Project Management Model

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

- Eric Peterson/Brett Johnson – Tula Foundation/Hakai Institute
- Mark Saunders/Stephanie Taylor/Caroline Graham – IYS/NPAFC
- Bruce Patten – DFO Pacific Biological Station and OBIS Canada Node Manager
- Gabrielle Canonico - GOOS BioEco Panel Co-Chair & NOAA Federal
- Evgeny Pakhomov - IYS Chief Scientist, Professor and Director UBC Institute of Oceans and Fisheries
- Dick Beamish and Brian Riddell – 2019 and 2020 Expedition Organizers
- National Chief Scientists for the IYS

For the 2022 expeditions to be successful, Data Mobilisation and Communication Plans should be developed with each 'National Lead Scientist' to develop and clarify the expectations related to data management. This approach should then be shared with Expedition Scientists in each Research Area of the High Seas Expeditions, with involvement spanning multiple nations. This should be done well in advance of the expeditions to identify every data element, method, platform, and variable they plan to collect as well as identify benchmarks of data mobility to aim for.

The IYS Theme Counsel Groups (TCG) provide a further mechanism for fostering broader international involvement in mobilising data generated through the IYS High Seas Expeditions. TCG 1: "Status of Salmon and Salmon in a Changing Salmosphere" should be responsible for consolidating Essential Salmon Variables that the IYS Data Mobilisation team can codify into canonical definitions hosted using a modern web-accessible ontology platform. TCG 2: "Human Dimensions" should help identify the major socio-cultural road blocks to mobilising salmon ocean ecology data as well as recommend how roadblocks could be removed first nationally and then internationally. TCG 3: "New Frontiers in Information Systems" should advise on the technical capacity and readiness of applying well-established new technologies to move salmon ocean ecology data management into the 21st century. TCG 4: "Outreach and Communications" should aid in the IYS Data Mobilisation team in translating technical and socio-cultural data management challenges into plain language for non-technical readers.

Science Model

GOOS provides a framework for globally integrated and sustained ocean observing. 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, and mobilisation. Through this scaffolding, data are regularly evaluated, and open data sharing is encouraged. Research Areas defined by the IYS should be mapped to the following GOOS science domains that reflect the Essential Ocean Variable (EOV) schema.

The following list indicates the relevant GOOS domains and three example EOVs that IYS

expedition participants ought to consider when defining their science model. Within each domain, Essential Ocean Variables should be mapped to the variables that will be collected by IYS scientists.

- Physics and Climate
  - Sea state
  - Ocean surface stress
  - Sea ice
- BioGeoChemical
  - Oxygen
  - Nutrients
  - Inorganic Carbon
- Biology and Ecosystem
  - Phytoplankton biomass and diversity
  - Zooplankton biomass and diversity
  - Fish abundance and distribution

Timeline

Three themes determine the time line:

1) Deepening Engagement and Impact; 2) System Integration and Delivery; and 3) Building for the Future.

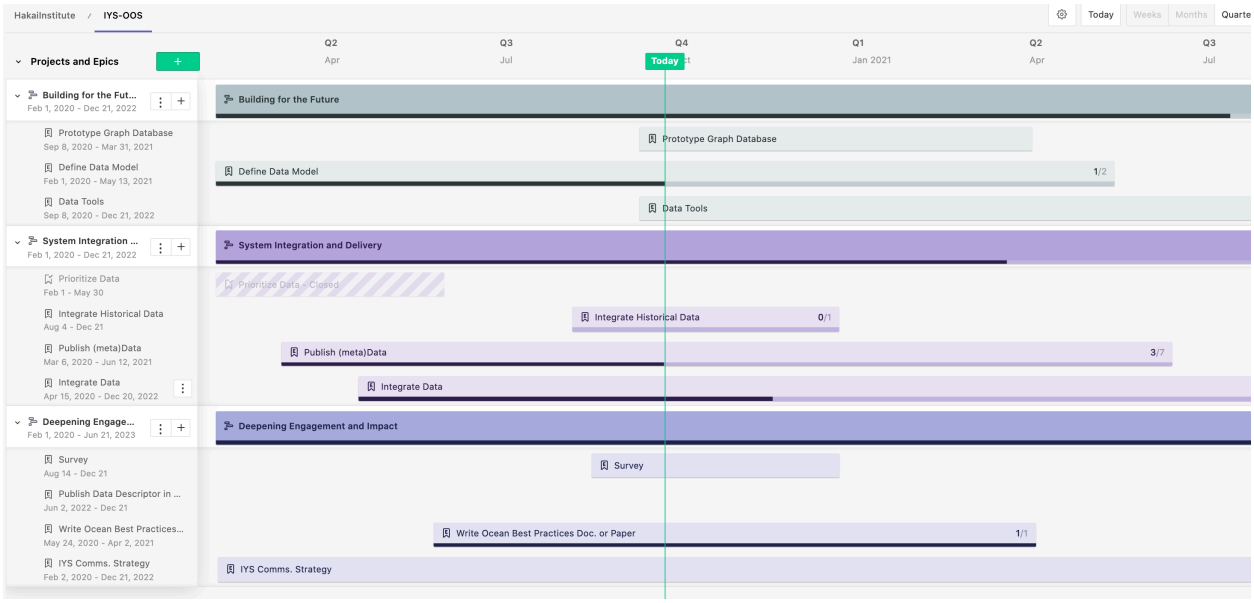

Figure 1. Data strategy road map through 2022

Next steps

Despite the challenges we face because of the uncertainty created by the CoVID-19 pandemic, we will remain focused on preparing for some form of coordinated research expeditions occurring in both 2021 and/or 2022.

- Continue to publish data sets provided by the 2019 and 2020 Scientists to the IYS Data Catalogue (https://iys.hakai.org).
- Transform and integrate 2019 and 2020 data sets to international standards and host data on appropriate repositories to demonstrate, only in novel examples/domains, how to achieve international integration.
- Facilitate Data Mobilisation and Communication plans for each IYS science group well in advance of 2021/22 Expeditions.
- Engineer research software starting with a GOOS-compatible technology kernel to facilitate programmatic data access, quality checks, visualisation, and other common data analysis functions.

Known Challenges

Across each IYS Research Areas, documentation of sensor or instrument calibration, collection protocols, sample processing protocols, data exchange protocols, and sample shipping and tracking protocols is critical. This includes equipment and platform descriptions, calibration files, instructions for how data were summarized or aggregated, and any caveats for data interpretation. This will help ensure the scientific integrity of the consolidated data sets. Dataset 'Quality' levels will eventually be assigned based on completeness of metadata collection and integrity of data provenance. Translators with domain specific knowledge of the data need to be identified in advance of data collection and integration. Development of best practices among research domains is pressing.

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. Using a change log to manually document changes to the raw data ensures that reproducibility can be achieved when data cleaning is performed ad hoc. Otherwise, we recommend moving towards automated version control, scripted data transformations and agreed upon controlled vocabularies and metadata profiles that canonically define variables. This will help scientists collaborate on common data sets and analyses using modern cyber-infrastructure that is already developed but require operational agreements to effectively implement.

Links and Resources

- IYS-OOS GitHub Repository
- IYS Data Catalogue
- Global Ocean Observing System

Contact us at iys.data@hakai.org for questions, comments, or data inquiries.