DRAFT SSWG Data Topics Recommendations
July 2022

Background

The scallop fishery is one of the most valuable fisheries in the U.S., generating hundreds of millions of dollars annually, but there is no dedicated NOAA funding or staff resources to ensure survey data standardization, access, and storage. The SSWG assessed the strengths and weaknesses of the current scallop survey data management system and determined that the system is inefficient, disjointed, and vulnerable to data loss and data merging errors. They noted that the system has been able to support science and management objectives to date, but emphasized that potential data loss, lack of coordination, or loss of experienced personnel could risk collapse of the data management process. The group concluded that the system is unsustainable in its current form. Currently, survey data products are stored on personal laptop computers, manually merged, and output in flat files with minimal ability for data sharing, accessibility, or repeatable analyses. The SSWG highlighted the critical need for dedicated survey data management personnel and database infrastructure.

Recommendations

1. **The Northeast Fisheries Science Center (NEFSC) should prioritize scallop survey data management and provide resources for dedicated personnel for data/database management.**

Rationale:

The "Foundations for Evidence-Based Policymaking Act of 2018" (Evidence Act; Public Law 115-435) requires that all NOAA data be open and usable by the public without restriction unless such sharing is expressly prohibited by law or regulation. NOAA's Data Strategy, released in 2020, outlines goals to align data management leadership roles across the organization, govern and manage data strategically, share data as openly and widely as possible, promote data quality improvements, and engage stakeholders to maximize the value of NOAA data (NOAA, 2020). These requirements and goals must be applied to scallop survey data management.

Implementation Strategies:

- The SSWG emphasized the need for the NEFSC to consider available and additional funding and staff resources to support scallop survey data management.
- The NEFSC should work with all scallop survey partners to identify methods to standardize data and increase efficiencies for survey data management.
- The NEFSC could consider prioritizing data needs as URGENT, IMPORTANT, and STRATEGIC to assess risk and vulnerabilities and inform contingencies for data storage, access, and delivery.

2. The Northeast Fisheries Science Center (NEFSC) should dedicate sufficient annual resources to develop and maintain an operational scallop survey data repository using FAIR (findable, accessible, interoperable, reusable) data management principles.

Rationale:
The SSWG highlighted that the current data storage approach is vulnerable to potential data losses, and the lack of data standardization can lead to data processing errors. Survey data merging and quality control is currently reliant on resource assessment specialists and is a burdensome, inefficient process. Scallop survey data are disjointedly housed by individual survey partners, and NOAA's current metadata portal (InPort: www.fisheries.noaa.gov/inport/) is not sufficient to support full data sharing of all sources of scallop data (e.g., NEFSC and RSA partners dredge and optical datasets.)

The FAIR data principles indicate that data should be findable, accessible, interoperable, and reusable (European Commission, 2018). The principles emphasize machine-actionability with machine-readable metadata for discovery of datasets.

- Findable: metadata and data should be easy to find for humans and computers
- Accessible: once found, users need to know how to access data
- Interoperable: data need to interoperate with applications for analysis, storage, and processing
- Reusable: metadata and data should be well-described so they can be replicated and combined

Implementation Strategies:

- The SSWG recommended that the NEFSC develop the scallop survey data repository to include standard data fields and quality assurance criteria that can be shared through web services in machine-readable format (e.g., JSON, XML, etc.).
- Initial development of the scallop survey data repository should focus on dredge survey data to inform database structure and identify integration and interface tools.
- The repository should be developed to allow additional survey data streams to be added and integrated.
- The NEFSC should explore cost and capability for storage of images from optical surveys.
- The repository must be operational beyond development phases and must be maintained in perpetuity.
- The SSWG recommended this as an URGENT priority to be initiated as soon as possible (Figure 1).

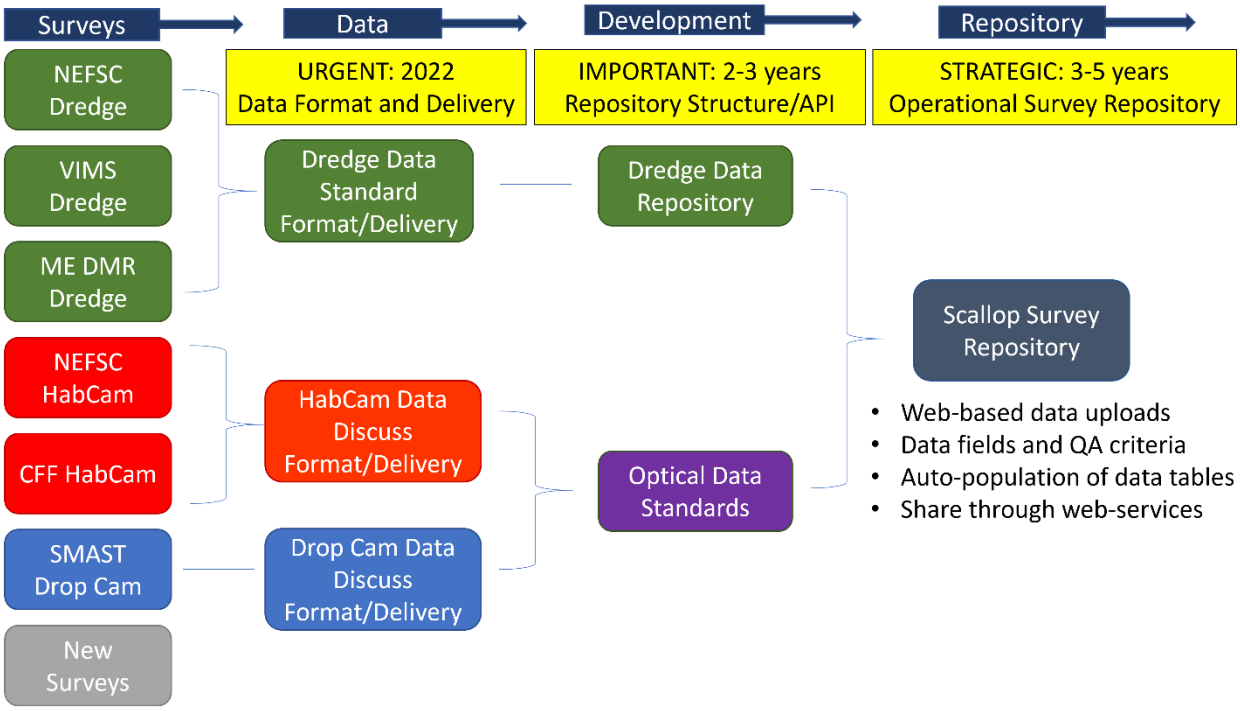

Figure 1. SSWG recommended steps and timeline for development of the scallop survey data repository.

3. Standardize scallop survey data format and delivery.

Rationale:

The SSWG noted that scallop survey data collection fields and protocols differ among the range of survey partners and that manual merging of datasets is time-consuming and error-prone. The SSWG identified standardization of survey data format and delivery to the NEFSC as a logical first-step to increase efficiencies in data management. Standardized data formats and delivery processes can serve as the basis for development of the scallop survey data repository.

Implementation Strategies:

- The SSWG recommended collaborations between the NEFSC and survey partners in the near-term (2022) to identify standard data fields and delivery format.
- Archived survey datasets should be standardized to facilitate integration in the scallop survey data repository within two to three years (by 2025).

4. Establish a process to check for autocorrelated data for model-based estimation methods.

Rationale:

Deriving biomass from geostatistical models is the preferred method for HabCam optical data. Low image annotation rates in areas with low scallop density have resulted in non-correlated data in some years, precluding the use of geostatistical modeling approaches to estimate biomass.

The SSWG recognized time and resource constraints for image processing and recommended an adaptive process to 1) identify whether or not data are autocorrelated and take steps to try to achieve autocorrelation, if possible; and 2) apply an alternative estimation method if autocorrelated data is lacking.

Implementation Strategies:

- The SSWG recommended that annotation rate expectations by survey/SAMS areas should be identified prior to the start of HabCam surveys.
- As soon as possible after image annotation, apply methods developed by the NEFSC to check for data autocorrelation prior to data delivery in a stepwise manner:
    1. Aggregate the annotated data by 750m segments
    2. Calculate Moran's I statistics for only the positive aggregated data points for each survey/SAMS area to check whether the data are spatially autocorrelated using reviewed methods (e.g., ArcGIS, QGIS, R function in Moran.I in library ape)
    3. If data are spatially autocorrelated ($p<0.05$), complete analysis and submit data
    4. If data are not spatially autocorrelated ($p>0.05$), review potential reasons for the lack of correlation with NEFSC and Council staff (e.g., too few images were annotated, or spatial structure is absent)
    5. In the absence of autocorrelation, the PDT will determine appropriate methods to generate biomass estimates (stratified mean estimation Chang et al., 2017).
- This process should be included in the Scallop Survey Guiding Principles and should be updated as needed.

5. Conduct a review of automated detection technology.

Rationale:

Manual annotation of optical survey images is resource-intensive and time-consuming. The annual scallop management process operates under a tight timeline and the preferred method to derive biomass estimates from HabCam data relies on relatively high annotation rates to support geostatistical models. Advancing the utility of automated detection technology could substantially increase the speed and rate of image annotation. Several scallop survey partners have developed automated detection tools, including training datasets, machine-learning algorithms, and detection software. The SSWG encouraged survey groups to continue development of automated detection and consider standardized training datasets to be used by all groups. The SSWG recommended that a review of the technology is needed in the near-term to advance the utility and application of automated detection.

Implementation Strategies:

- The NEFSC and Council should prioritize organization of a peer-review process to advance the utility of automated detection technology.
- Define objectives and Terms of Reference for a review of automated detection technology, including, but not limited to the following:
    - Identify what software has been applied and what tools are useful
    - Define data products and statistical analysis of accuracy and precision
    - Consider pathways to operationalize automated detection

- Identify an appropriate review panel with technical expertise, for example:
    - Regional Fisheries Science Centers
    - NOAA Center for Artificial Intelligence
    - NOAA Automated Image Analysis Strategic Initiative
    - ICES Working Group on Machine learning in Marine Science
- The SSWG recommended this an URGENT priority to be initiated as soon as possible. The review should include all relevant survey partners and not be conducted as part of a Research Track or Management Track Assessment.

6. The New England Fishery Management Council (NEFMC) should maintain data tables for management applications.

Rationale:
In 2021, the Council compiled survey data products in a single location, including survey biomass, projected exploitable biomass, and allocated and landed pounds by year, region, SAMS area, and survey type for 2015 to 2021. The compiled data facilitated analyses to support evaluation of rotational management performance, projection performance, and understanding of the impacts of various management measures. The SSWG noted that continued maintenance of the compiled survey data products would be useful for scallop science and management.

Implementation Strategies:
- Council staff should continue to review and update data tables on an annual basis.
- The Council should consider potential mechanisms to share data products and/or identify potential partners/services to house data with public accessibility.