Historical BIONESS Archival

Emily O'Grady

This R markdown outlines the steps taken to archive historical BIONESS data in BioChem. This project was undertaken under the direction of Catherine Johnson at Fisheries and Oceans Canada, Maritimes Region Ocean Ecosystems Science Division.

This report summarizes the completion of the project involving the consolidation and quality control of plankton data from various missions. The data was consolidated from various locations and shared folders at Bedford Institute of Oceanography from missions between 2009 and 2016.

Data cleaning and formatting

The raw data was consolidated into a single folder for each cruise, sourced from various locations. We checked the plankton data loaded to BioChem for each mission, noting that none included BIONESS data. Metadata was checked for each mission, and BIONESS log sheets were found for each mission since most missions required at least some metadata from the logs to be input.

A discussion with Rebecca Milne prompted the organization of data accounting by individual station, as data was often analyzed by location rather than mission. For example, Gully tows from 2003-2010 were analyzed together, likely due to funding organization. We reviewed log sheets and set up an accounting spreadsheet for 2009-2016 data by mission and station.

How many missions and stations were archived as part of this project

This project prepared and archived 55 stations across 13 missions from 2009-2016.

Table 1: Data archived under this project

Mission	Year	Number of stations
HUD2009005	2009	2
HUD2009048	2009	3
HUD2010006	2010	4
HUD2011004	2011	4
HUD2011043	2011	7
HUD2012042	2012	6
HUD2013004	2013	5
HUD2013037	2013	14
HUD2014004	2014	11
HUD2014030	2014	7
HUD2015004	2015	7
HUD2015030	2015	7
HUD2016003	2016	4
HUD2016027	2016	10

Metadata cleaning and supplementing

The process of metadata cleaning and supplementing was a crucial step in ensuring the integrity and usability of the plankton data for future analyses.

Initial Assessment and Consolidation

The first step involved consolidating raw data from various sources into a single folder for each cruise. This included data from Shelley Bond's investigative folders (from Mary Kennedy's archives) and SRC cruise folders. We then conducted an initial assessment of the metadata to identify gaps and inconsistencies.

Identifying Metadata Gaps

During the initial assessment, we identified several gaps and inconsistencies in the metadata. Common issues included missing sample IDs, sounding data, and discrepancies in date formats. Sounding data, typically recorded on BIONESS log sheets, was often missing and needed to be entered manually. Sample IDs were inconsistently recorded,and stickers were not used on some missions, requiring compound unique sample IDs to be generated. In those cases, the formula used to create compound sample IDs followed Mary Kennedy's procedures.

compound sample ID formula Example: 2013037092006002 2013037 is the numeric section of the mission identifier (HUD2013037) 092 is the last three digits of the gear type from BioChem bcgears table 006 is the Tow number 002 is the net number

Standardizing Metadata

To address some issues, we standardized the metadata across all files. This involved:

Ensuring consistent date formats (DD-MON-YY). Standardizing column names and metadata formatting, including dates and gear representation. Adding missing metadata manually from handwritten log sheets, such as event numbers, dates, and sample IDs.

Station names were standardized to facilitate future data extraction and comparison.

Table 2: Original and updated station names

Original Names	Updated Names
GULD3	GULLY_D3
GULLYD3	GULLY_D3
HL3	HL_03
HL_3	HL_03
RL1	RL_01
RL_1	RL_01
CSL_4	CSL_04
RATBA-1	RATBA_01
RATBA-2	RATBA_02
HL_3.3	HL_03.3

Quality Control and Validation

Quality control was a critical part of the metadata cleaning process. We conducted thorough checks to ensure:

- Dates matched between log sheets and data.
- Start and end depths were consistent between log sheets and data.
- Tow and sample numbers were accurate.
- Event numbers were consistent across log sheets and data.
- Results of these checks were recorded in a tracking spreadsheet, and any inconsistencies were documented in a metadata discrepancies file

Each inconsistency was investigated and resolved with help from data providers (Rebecca Milne) and supervisor (Catherine Johnson)

Supplementing Metadata

In cases where metadata was missing or incomplete, we supplemented it using various methods:

- Filling in sample IDs using a compound ID formula where necessary.
- Confirming metadata with team members, such as checking what was written on sample bottles when stickers were not used.

In some older missions (pre 2014), there were no elog or equivalent files. In these cases, elog tables were manually generated from handwritten bridge log and BIONESS log sheet information. These elog files match the format used to load elog metadata into the AZMP template.

Load file formats

The standardized format of the files is essential for ensuring consistency and ease of use in future analyses. Each file follows a specific structure with clearly defined columns. Below is an overview of the file format.

Standardized File Structure Each file contains the following columns:

Table 3: Data columns and definitions

Column Name	Definition
MISSION	Mission name
DATE	Date of collection
STN	Station name (standardized)
TOW#	Numeric identifier for the tow
GEAR	Mesh size in um (202)
EVENT	Numeric identifier for the event
SAMPLEID	Unique sample ID
START_DEPTH	Opening depth of the net
END_DEPTH	Closing depth of the net
ANALYSIS	Numeric code based on AZMP protocol
SPLIT	Numeric split
ALIQUOT	Aliquot of sample analyzed
SPLIT_FRACTION	Split fraction of sample analyzed
TAXA	Taxonomic name
NCODE	Code for taxonomic name to reference BioChem tables
STAGE	Numeric code for stage (reference to BioChem tables)
SEX	Numeric code for sex (reference to BioChem tables)
DATA_VALUE	Numeric measurement, either count or weight
DATA_QC_CODE	Quality control flag, following BioChem BCQUALCODES table
PROC_CODE	Numeric code based on AZMP protocol

Column Name	Definition
WHAT_WAS_IT	Numeric code based on AZMP protocol
COMMENT	Any comments on samples. This is also where we have input 'flag' comments

Date Format

The date format used in all files is DD-MON-YY to ensure consistency across datasets.

Extracting volume data from electronic files

In order to properly represent the plankton data in BioChem, the sampled volume information was essential. This metadata was found in the BIONESS electronic files.

The electronic file format was confirmed with Nelson Rice who wrote the original program to produce the files.

Volume data is combined with other metadata including event, and sample ID to be matched up to taxonomic data. This work was done in `volume_extraction.R` within this project.

Quality Control

Quality control functions were developed as part of this project. They can be found in the BIONESSQC package on GitHub (https://github.com/EOGrady21/BIONESSQC).

These tests were developed in close co-ordination with Rebecca Milne to meet the expectations of zooplankton analysis methodologies for the Atlantic Zone Monitoring Program at BIO.

1. Data Formatting Check: This function verifies the formatting of the BIONESS data, including column names, data types, and consistency in columns that should be copied. It checks for missing columns, correct data types, and consistency in specific columns like "MISSION" and "GEAR".

2. Metadata Check: This function compares the metadata (Elog file) against the BIONESS data to ensure that all events are present in both data frames and that the date ranges match. It checks for missing events in the metadata and mismatched date ranges. This test is optional as some missions did not have an elog file generated.

3. Calanus Analysis Check: This function ensures that all Calanus species are labeled as analysis 2 and have the same split fraction for each sample ID. It checks for consistent analysis values and split fractions among Calanus species.

4. Weight Checks: This function verifies the weight columns in the BIONESS data. It ensures that the dry weight is less than half of the small wet weight for a unique sample ID, stage and sex, and that the total weight is consistent with the sum of the small and large biomass. It also checks for properly named weight data.

5. Data Completeness Check: This function checks the BIONESS data for completeness, including the presence of multiple zooplankton species in analysis 1, the presence of small biomass, total weight, and dry weight, and the presence of Oithona species. It verifies that specific values are correctly coded for small biomass, total weight, and dry weight.

6. Split Fraction Check: This function ensures that wet weights have a split fraction of 1, dry weights have a split fraction of 0.5, and that Calanus species have a single unique split fraction for each sample ID. It checks for correct split fractions in the data.

7. Large Animal Check: This function verifies the large animal counts and weights to ensure that there is only one count per sample ID and that the counts are integers. It also checks that there is only one weight per sample ID and that the weights are recorded to 4 decimal places.

8. Numeric Precision Check: This function ensures that all counts are integers and all weights are recorded to 4 decimal places. It checks for non-integer counts and integer weights in the data.

Quality Control Results

The column DATA_QC_CODE was used to represent the quality control results in the raw data sheets for each mission. It has been requested by OESD and ODIS to add this column to the BioChem plankton table structure, in BCPLANKTNGENERALS. As of the date of this report this change has not yet been made, and so the BIONESS data has not yet been loaded to BioChem.

The flags in DATA_QC_CODE follow the BioChem flag scheme.

0 - QC has not been performed.

1 - QC has been performed; element appears correct.

2 - QC has been performed; element appears inconsistent.

3 - QC has been performed; element appears doubtful.

4 - QC has been performed; element appears erroneous.

There were no major quality control issues detected in the historical datasets in this project. There were a handful of data entry errors and formatting issues that were caught and corrected. There were also a handful of cases where unique circumstances (ie. broken nets) caused slightly skewed results, and so we were able to apply flags to these data, in combination with comment text in order to inform the end-user.

Conclusion

This project was able to successfully archive a significant chunk of historical BIONESS data, although more data remains. The missions which were processed under this project have all been archived on BIO's shared drive, under the appropriate mission folder in the general SRC decade folders. Although the data has not yet been loaded to BioChem, it is the intention that this project continue and the data be loaded through DART once the required adjustments are made to BCPLANKTNGENERALS.

Continuing work

Patrick Upson and Emily O'Grady are continuing to work on the capacity of DART to load stratified net data through DART to BioChem.

There are approximately 43 more stations (tows) of BIONESS data to be archived. Some of this data has not been analyzed and will require future effort from zooplankton taxonomists and data management support.

An inventory of all BIONESS and multinet data is available on sharepoint (link below in **Resources**) in the sheet: "Stratified_Net_Data_Accounting_2009_2024". Which can help to direct future efforts.

Resources

The details of this project are contained in sharepoint

The BIONESSQC package can be found on GitHub at link

Other BIONESS workflows can be found in this project on GitHub link

Individual mission files including BIONESS electronic files, QC results, and taxonomic data can be found on SRC under the appropriate decade and mission folders in a standard 'BIONESS' folder structure. eg. "R:/Science/BIODataSvc/SRC/2010s/2014/HUD2014030/BIONESS"