Shoals Marine Laboratory
Long-term Rocky Intertidal Monitoring Program
Documentation of data quality
An T Nguyen, 2018 Environmental Data Intern
August 2018

The rocky intertidal zone around Appledore Island, ME, is home to a wide array of biodiversity. A program was established here in the 1970s to monitor the presence and abundance of key intertidal seaweeds and invertebrates. Monitoring was originally conducted by students enrolled in courses at the Shoals Marine Laboratory (SML), but the program has since evolved into a 2 ½ week internship at SML. All combined, the dataset documents almost 40 years of intertidal biodiversity with approximately 150 unique taxa identified. There are few comparable monitoring projects in the Gulf of Maine. Thus, this dataset is valuable for determining patterns and changes in species composition and distribution in the rocky intertidal habitat, especially in the context of a rapidly changing and ecologically important region. However, the program's diverse history and data management practices has hitherto impeded data discovery and utilization.

This dataset underwent an intensive cleanup effort in the summer of 2018 by An T Nguyen with support from Drs. Kylla Benes, David Buck, James Coyer, Chris Siddon, and Hal Weeks, funding and mentorship from Shoals Marine Laboratory and the Environmental Data Initiative. This document lists the issues that were encountered in this period and the methods used to resolve them. Most manipulations were conducted using the statistical software package R, utilizing RStudio and the tidyverse libraries. **We recommend that anyone planning to make use of the data read this document closely after the accompanying protocol to become aware of those issues and implement appropriate measures.**

Finally, please contact Shoals Marine Laboratory to tell us how you are using this data so that we can track usage. Please send an email describing use to shoals.lab@unh.edu. We thank you for your time in reaching out to us. Our website: shoalsmarinelaboratory.org.

1. General data organization

1.1 Data files and worksheets

Originally, the dataset consisted of about 300 Excel spreadsheets, each containing data from one transect site in one year. Each Excel file contains at least four worksheets in this order: Percent cover, Size, Counts, and Categories. Details about how each of these

subsets of observations were obtained can be found in the ‘SML Intertidal Monitoring Program’ document. Since each worksheet corresponds to a different quantification method, we have chosen to work with them separately despite some taxonomic overlap. Thereafter in this document, when there is reference to “percent cover,” “counts,” or “categories” these denote those different subsets of the larger data set.
Important note: we found that Categories as a dataset is not very reliable because different methods of quantification were used over time. We recommend that data users use this subset for presence/absence data only.

Each worksheet has several "ID" columns denoting the year, transect number, level below transect pin (a proxy for tidal height), and the replicate quadrat number (if implemented), then anywhere from a dozen to fifty additional columns of species records. Thus, each row contains information on one single quadrat.

1.2. Taxonomic reconciliation

In the process of merging datasheets together, we found that:
- taxa were documented under abbreviations; these abbreviations often differed among years, resulting in the same organism documented many different ways
- sometimes taxa were identified to different resolutions (i.e., species, genus, order, or higher)
- not the same group of organisms were present as column headers in every year; on the other hand, some templates were copied from year to year and not changed, and subsequently some columns containing uncommon taxa were left empty
- there were discrepancies in the way macroalgae were recorded under primary/canopy cover (more on this in a later section)
- since the program was driven primarily by students, some taxa were emphasized and quantified only in some years based on student interest. For example, most worm species were not quantified after 2004

In most cases we were able to infer the appropriate taxonomic name of each organism. If there was any concern, those are documented in the accompanying species list. We follow the World Register of Marine Species (WORMS) when it comes to taxonomic authority, and the taxonomic information included is from this source.

1.3. Canopy/Primary distinction

Some macroalgae species form a distinct "canopy" when submerged at high tide; this canopy is separate from the “primary” organism cover that is directly attached to the rock substrate (see illustration in protocol). Primary cover can include smaller macroalgae, macroalgal holdfasts, and sessile invertebrates. The percent cover of a number of macroalgae are split into two columns; e.g., "Asco primary" and "Asco canopy" referring to the primary and canopy cover of *Ascophyllum nodosum*, respectively. During our data organization, we found some species did not consistently receive this treatment. In some years the primary/canopy distinction was made (e.g., "Ceramium primary" and "Ceramium canopy" columns present in the datasheet) and in other years it was not (e.g., there is only one column titled "Ceramium").

Where this happened, we chose to either 1) unite all entries under the general taxon name (e.g., "Ceramium") and discard the primary/canopy distinction or 2) assign either a primary/canopy distinction to the generic entries. Solution 1, of course, would create situations where there are two entries of "Ceramium" in the same quadrat. We have chosen to keep the higher of the two entries in these cases because, for these species, most observations would be the total amount of space occupied as visualized from above the quadrat, analogous to canopy cover.

A list of macroalgae species where this was a issues and their respective treatment (either 1-discard or 2-assign):
Acrosiphonia: discard
Ahnfeltia plicata: discard
Ceramium virgatum: assign to canopy
Cladophora: discard
Colpomenia peregrina: discard
Corallina officinalis: discard
Dumontia: discard
Leathesia marina: discard
Polysiphonia: assign to canopy
Porphyra: discard
Rhizoclonium: discard
Ulva/Ulva lactuca: discard
Vertebrata: discard
Pylaiella littoralis: discard
Elachista fucicola: discard

1.4. Uncertain taxonomy
These include cases where only a higher classification (i.e., genus) was named or only a portion of the name was documented. When possible we renamed the entries to be more specific. Considerations during this process include whether there is only one species of that genus in the Gulf of Maine, and there being no reports of another (e.g., "Asco" to "Ascophyllum nodosum") and knowledge of historic naming schemes. For example,

while “Ulva” is the genus of several green seaweed species in the area, historically “Ulva” was synonymous with “Ulva lactuca”; whereas, "Enteromorpha" referred to "Ulva intestinalis." Uncertain taxonomies that could not be rectified were left as is; typically to the genus or higher order classification level.

2. Data organization and formatting

2.1. Worksheet formats

The "wide" format originally adopted by SML is common in community datasets, ideal for data entry, and for obtaining an overview of the quadrat. However, wide formats are not as conducive to merges and manipulations as the "long/tidy" data format as archived in the data package. It is not recommended to turn the datatables back into wide format; doing so will introduce new columns where there were none in the original datasheets; sometimes NA values would be converted into zeros as well.

2.2. Missing values

Most cells in the original datasheets are blank—a common occurrence in community ecology data since most species are uncommon and it saves time in data entry. However, it is important to ascertain if any given blank cell is actually a zero value (indicating that data for that quadrat was taken, but none of that organism was observed) or a "NA" value (data for that quadrat was not taken due to time constraints, dangerous wave action, etc.). Starting from 2009, a new column "Data_taken" was added to all datasheets; variations on "yes/no" in that column indicate whether data on that quadrat was observed or not. However, there is no such information available for data prior to 2009.

We chose to retroactively determine if a blank cell is a zero value or a NA value based on the following set of criteria: 1) where available, "Data_taken" is y/n, 2) if there is data elsewhere in the quadrat, 3) if there is data on percent cover of organisms and/or bare rock of the quadrat, since this subset is most likely to have data. Additionally, we retroactively filled in y/n values in the "Data_taken" column where there were none, using the same criteria.

2.3. Non-quantifiable values

If students left notes on the presence of certain organisms that are not quantifiable ("many," "few," etc), then, at data entry these were noted as either "nd" for "no data," or "p" for "present." This happened more often in the earlier years of the program. We chose to convert all "nd" values to "p" to avoid confusion. Note: "p" values were used when determining the value of blank cells (see 2.2. Missing values); that is, when a quadrat has nothing but "p" entries, the other blank cells were determined to be 0.

2.4. Non-numeric values
Where data cells contain non-numeric characters (e.g., 15% or ~400), these characters were examined manually and removed if appropriate.

3. Changes in data entry

3.1. Filename issues
Occasionally a data entry template is reused in subsequent years and the ID information within the datasheet does not get updated, hence it does not match the filename or the enclosing folder. We inspected these data and compared them to the same transect from previous years and other transects from the same year to make sure they indeed corresponded to the filename.

3.2. Quadrat coding
Prior to 2009, transect number, level from transect pin, and replicate number if applicable, were not recorded as separate columns but combined into "quadcodes." For example, a quadcode of "t15l4r1" would translate to "transect 15, level 4, replicate 1." We extracted the appropriate values from quadcodes and placed them into the appropriate columns. When they did not make sense, original datasheets were inspected manually and edited.

Occasionally data above the transect pin is taken; this happens when there are organisms above the pin and surveying continues landward until there is nothing but bare rock. These levels are mostly denoted as negative values; e.g., level -1 is 1ft in height above the pin, or 14.5ft above MLLW. There were a couple instances where these were denoted as 99 or 98 instead; these were changed to -1 and -2 respectively.

3.3. Size class data
There are six species (*Ascophyllum nodosum, Fucus spp., Semibalanus balanoides, Mytilus edulis, Modiolus modiolus, Nucella lapillus*) that are documented in the "Sizes" tabs in most worksheets.

For the invertebrate taxa, prior to 2011, the prevalent practice was to record the total count of these species in the "Count" worksheets, and then to divide the relative abundance into different size classes (i.e., total abundance across size classes was 100%). For example, 300 *Semibalanus balanoides* in "Counts" worksheet, would be split into 70 (%) 0-2mm and 30 (%) 3-5mm sized barnacles in the "Sizes" worksheet. In 2011 and thereafter, this practice was changed to record the absolute count of each species in each size classes. For example, the same data as above would be recorded as 300 barnacles in "counts," 210 0-2mm and 90 3-5mm in "sizes."

We chose to follow the post-2011 convention and retroactively calculated the absolute counts in each size class for the pre-2011 data, using the respective percentages multiplied by the total counts. Exceptions are when the total counts are 0 and only percentages are available. In these cases we filled "p" for all associated entries to avoid misdirection. On the other hand, occasionally in post-2011 data there are absolute counts in each size class, but no corresponding total record of the species in the "Counts" worksheet. We corrected these cases as well.

Prior to 2008, *Mytilus edulis* and *Modiolus modiolus* each were documented under five size classes (0-5, 6-10, 11-20, 21-30, and >30mm). After 2008, a sixth size class, >50mm, was added to each species. There were rarely any mussels of this size in the dataset, so we decided to drop this class going forward, and put any that was recorded with the >30mm bin.

We chose to extract the information on *Fucus* and *Ascophyllum* to a separate datatable, fucus_asco_max_size.csv.

4. Notes

The original datasheets contained copious amounts of comments and notes. Students and interns were instructed to put down information on any organism without its own column in the comments. Sometimes these notes contain substantial information (e.g., 40% of percent cover). Data entry personnel occasionally commented on the “quality” of the data in the quadrat as well.

It was not feasible in the summer of 2018 to parse these notes; they are compiled as a separate datatable, notes.csv, and included in the data package. Note that there could be multiple notes/comments pertaining to the same quadrat.