A “data pipeline” for BIOS-SCOPE

Poster by Krista Longnecker

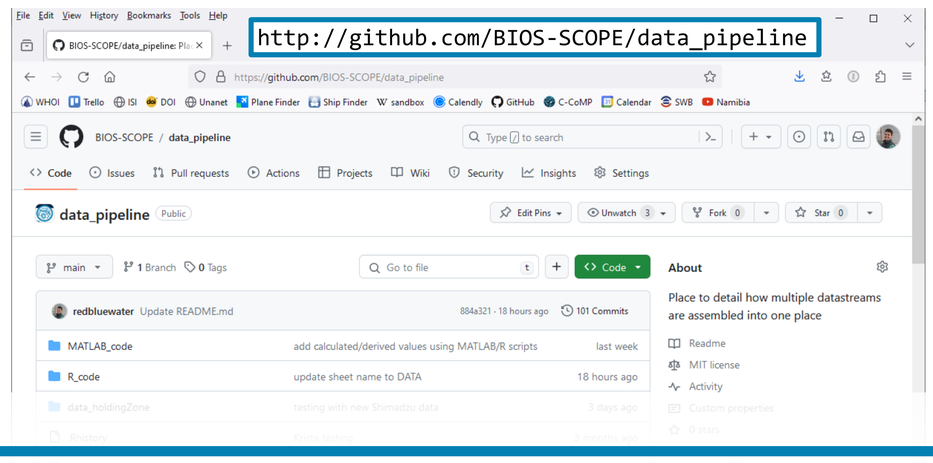
Pipe line is based on code from Shuting Liu and Ruth Curry

With thanks to Elisa Halewood and Rachel Parsons

What is the goal?

We want a single file that combines CTD data, variables calculated from CTD data, and data from discrete samples.

All of this code is available online at GitHub



Step-by-step:

BATS team: Processes CTD data and uploads to DropBox

↓

Rachel: Moves files to BIOS-SCOPE’s Google Drive where they are accessible to anyone in BIOS-SCOPE

↓

In R: Add new CTD data to master file using `Join_BATS_All_with_master_v3.R`, code originally from Shuting

↓

In R: Add discrete data to master file (including the derived variables calculated in MATLAB) with `Join_discreteData_v2.R`

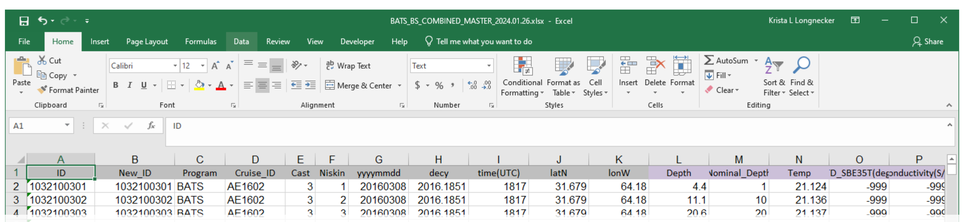
↓

End result: Excel file `BATS_BS_COMBINED_MASTER_2024.01.26.xlsx` available to anyone in BIOS-SCOPE (sits in Google Drive at: `1.0 DATA / 1.0 CURRENT BOTTLE FILE/`)

In MATLAB: Calculate derived variables from CTD data using Ruth’s code that is now `create_bioscope_files_2022_2023_Krista.m`

Other discrete data: Can easily be merged in if New_ID is one column.

BATS_BS_COMBINED_MASTER_{date}.xlsx



As of January 2024: 18,071 rows of samples and 137 columns of variables

Another way of describing this pipeline:

at sea	*Generate CTD data*		*Collect water for discrete samples*
on land	BATS team: CTD data processing		Process discrete samples (in lab)
	BATS team: move data to DropBox		End result: data file for each set of discrete samples
	Transfer data to BIOS-SCOPE Google Drive		
		Assigned to: Rachel Parsons (or Craig Carlson)	
	Get R script from github.com/BIOS-SCOPE/data_pipeline/R_code		
		Join_BATS_All_with_master_v3.R	
		Status: opens up Excel file with new data in rows	
		Next step: append the new rows to the existing bottle file; update the log; save Excel file with new date	
	Get MATLAB scripts from github.com/BIOS-SCOPE/data_pipeline/MATLAB_code		
		Create_bioscope_files_2022_2023_Krista.m	
		Status: generates CSV file with the calculated variables	
		Next step: use Join_discreteData_v2.R to pull the variables into the bottle file	
	Get R script from github.com/BIOS-SCOPE/data_pipeline/R_code		
		Join_discreteData_v2.R	
		Status: generates one Excel sheet with all the discrete data	
		Next step: copy the new worksheet into the discrete file, copy the header with the colors, update the log	

Acknowledgements:

This work is supported by funding from the Simons Foundation International’s BIOS-SCOPE program.