Helpful tools for accessible and reproducible research

Hansen Johnson

PhD Student Oceanography Department, Dalhousie University hansen.johnson@dal.ca

> MEOPAR Annual Training Meeting Victoria, BC June 11, 2019

Presentation online at:

https://hansenjohnson.org/talk/2019_meopar_atm/

Writing 00000000 Documentation

Tools of the trade

Many academic programs teach research concepts, but expect technical skills

- Most projects rely heavily on technology
- Little time or resources are allocated to developing technical skills and best practices
- These can make a HUGE impact for accessibility, efficiency and reproducibility
- Students must spend their limited time learning for themselves

Analysis

Writing 00000000 Documentation

My background

- Biology major in undergrad
- No training in computer programming or technical aspects of research before starting grad school (2015)
- Given a project that is impossible without technical chops





Writing 00000000 Documentation

My background

- Luckily I have interest and supportive advisers
- Developed many helpful skills with help from my peers (especially Christoph Renkl) and the internet
- Hope to help others acquire these skills more efficiently

Some examples:

Methods in Ten Minutes: https://christophrenkl.github.io/mtm/

R/Python Programming Tutorials: https://christophrenkl.github.io/programming_tutorials/



Writing 00000000 Documentation

Today's Goal

Goal: Provide some tools and concepts that I find essential for research

- Imagine we've been given some data on sea ice coverage and asked to characterize how it has changed over time
- Approach this simple project in 3 steps:
 - Analyse the data
 - Write a report
 - Occument the workflow
- We'll pause briefly after each section for questions and/or discussion

Disclaimer: these are the subjective opinions of a non-expert

Analysis •••••••••• Writing 00000000 Documentation

Analysis

Goal: Process and plot some data

- Structure the project
- Pead, process, and save data
- Make and save plots

You will need

R (www.r-project.org) Rstudio (www.rstudio.com)

Analysis

Writing 00000000 Documentation

A good project structure

A well-structured project allows you or someone else to easily understand and even reproduce the workflow

Organizing a project helps you:

- Expand, revisit and update efficiently
- Have confidence in the results
- Collaborate easily

Writing 00000000

Project structure

Projects vary and organizing them is hard. Some tips:

- Keep an untouchable 'sacred data directory' for raw data
- Dedicated directories for outputs (processed data and plots)
- Use simple file/folder names (ideally without spaces)
- Try to be consistent among projects
- Document prolifically (more later)

Check out CookieCutterDataScience for more details

 otivation
 Analysis

 000
 000

Writing

Documentation

Example [simple] project structure

example	Project directory
	All data
processed	. Processed data by code in src
	Raw data - never touch!
figures	. Plots produced by code in src
reports	Any reports or presentations
src	All source code
wrk	Development sandbox
readme.md	Project description
master.R	Master script

Analysis

Writing 00000000 Documentation

R and Rstudio



The basics of R and Rstudio are outside the scope of this session. See the tutorial here for more information: https://christophrenkl.github.io/programming_tutorials/

Analysis

Writing 00000000 Documentation

R and Rstudio

- Open Rstudio
- Create new project in a logical place with a short, descriptive name (e.g., ~/Projects/ice_cover)

New Project		M reports /2010_0
Back	Create New Project	
	Directory name: ice_cover	
K	Create project as subdirectory of:	
	~/Projects	Browse
	Create a git repository	
	Use packrat with this project	
c		
i 🗌 Open in new s	ession Cr	eate Project Cancel

Analysis

Writing 00000000 Documentation

Get the data

Download data from: https://www.canada.ca/en/environment-climate-change/ services/environmental-indicators/sea-ice.html

Save the file in data/raw/

Motivation Analysis 0000 000000€0

Analysis ○○○○○○●○○○○○○○○ Writing 00000000 Documentation

Process the data

Create a script called src/process_data.R to:

- Read in data from data/raw/
- Clean and format
- Save output in data/processed/

Analysis

Writing 00000000 Documentation

src/process_data.R

```
## process_data ##
# Read, process, and save ice cover timeseries data
# input -----
# choose data file
infile = "data/raw/1.SeaIce-NCW-EN.csv"
# choose output file
outfile = "data/processed/ice_cover.rda"
# process -----
# read in data and rename columns
df = read.csv(infile, skip = 2, col.names = c("year", "ice_cover"))
# remove missing values
df = df[complete.cases(df),]
# format year
df$year = as.numeric(as.character(df$year))
# save
save(df, file = outfile)
```

Analysis

Writing 00000000 Documentation

Plot the data

Create a script called src/plot_timeseries.R to:

- Read in data from data/processed/
- 2 Make plot
- Save output in figures/timeseries.png

Analysis

Writing 00000000 Documentation

src/plot_timeseries.R

```
## plot timeseries ##
# Make and save an ice cover timeseries plot
# input -----
# data file
infile = "data/processed/ice_cover.rda"
# plot file
outfile = "figures/timeseries.png"
# setup -
# external libraries
library(ggplot2)
# process -----
# plot
plt = ggplot(df)+
 geom_path(aes(x=year, y=ice_cover))+
 labs(x="Year", y=expression(paste("Sea ice area [million"," km"^"2","]")))+
 theme_bw()
# save
ggsave(plt, filename = outfile, height = 3, width = 5, units = "in", dpi = 300)
```

Analysis

Writing

Documentation

figures/timeseries.png



Analysis

Writing 00000000 Documentation

Simple project orchestration with a master script

Create a master file to execute all the analysis steps in the correct order. This should:

- Run src/process_data.R
- 8 Run src/plot_timeseries.R

Analysis

Writing 00000000 Documentation

master.R

master
Process and plot example ice cover timeseries
process raw data
source("src/process_data.R")
plot timeseries
source("src/plot_timeseries.R")

Analysis

Writing 00000000 Documentation



The project is totally reproducible from raw data! Now you can:

- Make changes to either the plotting or the processing script
- Delete anything in data/processed Or figures

And simply run master. R to re-build the entire project!

Analysis

Writing 00000000 Documentation

Key concepts

- Never edit raw data!
- All processed data and figures should be reproducible from raw data
- Use a master script (or other means) to orchestrate data processing
- Take time to improve code readability (use comments, indent, consolidate inputs, etc.)

Possible next steps

- Use Make instead of a master script to orchestrate the project more efficiently
- Use symlinks to link to large datasets that are stored remotely
- Use functions for repeated tasks

Writing 00000000 Documentation

BREAK

Questions?

How do you keep your projects organized?

Analysis

Writing •••••• Documentation

Writing

Goal: Find and organize references and draft a research report

- Find references
- Organize and review references with Zotero
- Write and cite document with Word / LibreOffice

You will need

Zotero (www.zotero.org)

LibreOffice (www.libreoffice.org) OR Microsoft Office [paid] (https://products.office.com/)

Analysis

Writing ••••••• Documentation

Introducing Zotero



An open-source, one stop shop for acquiring, organizing, reviewing, and citing references

Motivation
0000

Analysis

Writing

Documentation

Acquiring

- Install Zotero plugin for web browser
- Pind a reference (usually w/ Google Scholar)
- Navigate to the journal page
- Right click anywhere on the page and select Save to Zotero (Embedded Metadata)

SHARE





Perspectives on the Arctic's Shrinking Sea-Ice Cover



Mark C. Serreze^{1,*}, Marika M. Holland², Julienne Stroeve¹

Science, 16 Mar 2007: Back Forward Reload	3-1536 26			
Save As Print Cast Translate to English	ures & Data	Info & Metrics	eLetters	🔁 PDF
Save to Zotero	Save to Zotero (HighWire 2.0)		
View Page Source Inspect	Save to Zotero (Save to Zotero (Save to Zotero (Embedded Metadata) DOI) Web Page with Snapshot)	ability in the coupl	ry month. ed ice-
Speech	Save to Zotero (Web Page without Snapshot)	orted by evidence o	f

Motivation
0000

Analysis

Writing

Documentation

Organizing

Open Zotero application and browse references. You can:

- Search / sort by author, year, journal, etc.
- Organize into project folders / collections / tags
- Add items from scratch

٠	• •		Zotero							
	s 🝙 · 🖉 🔏 🔍 / ·	🔍 Q.	sea ice co	ver	0	• •				¢
Tit	e	Publication	Creator	Year	D.	Info	Notes	Tags	Related	
•	Thinning of the Arctic sea-ice cover	Geophysic	Rothroc	1999			- Notice	lugo	nonatou	
	🔁 Rothrock et al 1999 - Thinning of th					Item Type	Journal A	rticle		
l ►	Subarctic cetaceans in the southern Chu	Oceanogra	Clarke e	2013		Title	Thinning	of the Arc	tic sea-ice	cover
►	Age and growth estimates of bowhead w	Canadian J	George	1998		✓ Author	Rothrock	D. A.	1	•
►	Comparing marine mammal acoustic ha	Polar Biology	Moore e	2011		- Author	Yu, Y.		1	•
►	Updated 1978-2001 abundance estimat	Journal of	Zeh and	2004		- Author	Maykut, (G. A.	1	• •
►	Satellite Tracking of Western Arctic Bow	Alaska Dep	Quaken	2010		Abstract	-			
►	Abundance and Population Trend (1978	Marine Ma	George	2003		Publication	Geophysi	cal Rese	arch Letters	
►	Assessing the potential of autonomous s	Methods in	Suberg	2014		Volume	26			
►	An Overview of Fixed Passive Acoustic	Oceanogra	Mellinge	2006		Issue	23			
•	Acoustically Detected Year-Round Prese	Conservati	Morano	2012		Pages	3469-347	2		
►	Relationship between the distribution of	Polar Biology	Murase	2002		Date	1999-12-	01		y m d

Analysis

Writing

Documentation

Reviewing

You can:

- View PDFs (with default viewer)
- Add notes / other files / etc
- Update / edit metadata
- Click and drag to share reference

	Z	otero		
	😹 * 🖉 * 🔍 🔍 Q* sea	ice cover 🛛 🔘	ф т	C
Title Thinning of the Arctic sea-	Publication Cre ce cover Geophysic Roti	ator Year 🕫 hroc 1999	Inte Notes Tags Related	
E Facts about ice cover!		н	tem Type Journal Article	
🗧 😑 😑 🔹 Rothrock et al 1999 - Thinning of the Are	ctic sea-ice cover.pdf (page 1 of 4	.) ~	Facts about ice	cover!
	· 👌 🖻 🔍 🤉	Search	BIUS×,× ² <u>A</u> × <u>A</u>	- <u>I</u> x 66 ⊗
			🖡 Paragraph 👻 🚍 🚍 🗄	= = H
GEOHIYSICAL RESEARCH LETTERS, VOL. 26, NO. Thinning of the Arctic Sea-Ice Cover D.A. Rothrock, Y. Yu, and G.A. Maykut University of Wahington, Smith, Wahington Datract. Comparison of mexics draft data acquirted In	23, PAGES 3469-3472, DECEMBER the 1990s many ice draft data have	, 1999 been acquired	A Facts about ice coverl	
on submarine cruises between 1993 and 1997 with simi- by t lar data acquircle between 1958 and 1976 ridincizes that which the mean ice draft at the end of the melt season has mari- decreased by about 1.3 m in most of the deep water data portion of the Arctic Ocean, from 3.1 m in 1988-1976 (US to 1.8 m in the 1990s. The decreases is greater in the and	he Scientific Ice Expeditions (SCIC) h provided the opportunity to use U nes for Arctic research [Gossett, 199] used here (Figure 1) are from Ses § Pargo), September-October 1996 September 1997 (USS Archerfish).	EX) program, J.S. Navy sub- b6]. The 1990s ptember 1993 i (USS <i>Pogy</i>), Data from the	Parent Item: Thinning of the Arctic se Related: [click here] Tags: [click here] URL http://doi.wiley.com/10.1029/19	a-ice cover

Analysis

Writing

Documentation

Write and cite

In Word / Libre:

- Install Zotero plugin
- Olick Zotero tab
- Add references and bibliography with desired style

The past 30 years of sea ice cover in Canada

June 11, 2019 Hansen Johnson

Sea ice has been in decline for many years (Rothrock et al. 1999). Stroeve et al., (2008) suggest it declined sharply in 2007. This has been confirmed by modeling efforts (Saucier et al. 2003, 2004). Figure 1 shows the timeseries. Here's another citation from {Citation}

Z • george

My Library

Abundance and Population Trend (1978-2001) of Western Arctic Bowhead Whales Surveyed Near Barrow, Alaska George et al. (2003), Marine Mammal Science, 20(4), 755-773.

Brief overview of the 2010 and 2011 bowhead whale abundance surveys near Point Barrow, Alaska George et al. (2011), Paper SC/64/AWMP7 presented to the IWC Scientific Committee.

Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization George et al. (1998), Canadian Journal of Zoology, 77(4), 571-580.

Observations on the ice-breaking and ice navigation behavior of migrating bowhead whales (Balaena mysticetus) near Point Barrow, Alask... George et al. (1988), Arctic, 42, 24-30.

Analysis

Writing

Documentation

Key concepts

Use Zotero to acquire, organize, review, and cite references

Possible next steps

- Use LATEX for writing reports
- Use LATEX beamer for making presentations
- Combine text, code and output into documents (html, pdf, word) and presentations (pdf, ppt, html) with Rmarkdown

Analysis

Writing

Documentation

BREAK

Questions?

What other tools do you rely on for writing?

Analysis

Writing 00000000

Documentation

Goal: Document your work so that you can easily revisit, revert, and share

- Add a readme file
- Iracking changes with git and Rstudio
- Remote backups and hosting with GitHub

You will need

git (www.git-scm.com) GitHub account (www.github.com)

Analysis

Writing 00000000 Documentation

Readme files

What is a readme file?

- Usually simple text (*.txt) or markdown (*.md) file
- Includes any information required to implement or interpret the project workflow

Common things to include:

- Brief project background (goals, motivation etc.)
- Description of contents
- System requirements (code, software, etc.)
- Any caveats or known errors / bugs
- To do list
- Links for more information

Analysis

Writing 00000000 Documentation

readme.md

```
# README
Simple project to provide examples of helpful tools and
concepts for efficient and reproducible research
## Goal
Review recent trends in Canadian sea ice cover
## Dataset
Sea ice cover data were downloaded here:
https://www.canada.ca/en/environment-climate-change/services/environmental-indi
## Contents
'data' - all data
  'processed' - cleaned and formatted data ready
  'raw' - only raw data *never touch*
'src' - R code
'wrk' - development sandbox
'reports' - all presentations, reports, etc
'figures' - all figures
'master.R' - master script to reproduce full analysis
'readme.md' - this file
```

Analysis

Writing 00000000 Documentation





- Git is a hugely popular version control system (VCS)
- Open source software designed to help you track and document changes to projects
- Originally designed to be run on command line, but many more convenient interfaces now (e.g., Rstudio)

Writing 00000000 Documentation

How does git work?

- git provides a convenient way to save a 'snapshot' of your project at a point in time
- Allows you to review project history and revert one or more files to a previous version
- You must add ('commit') changes to one or more files to the project timeline, and provide a description of your changes



MotivationAnal0000000

Analysis

Writing 00000000 Documentation

Using git in Rstudio

- Navigate to Tools -> Version Control -> Project Options -> Git/SVN and switch Version Control System to Git
- Prestart Restudio

	~/Projects/ice_cover - RStudio	
?	Confirm New Git Repository Do you want to initialize a new git repository for this project?	Connection aset - 🛷
Project Op	No Yes	
R General	Version control system: Git	
Code Editing	Origin: None	
Sweave Sweave	⑦ Using Version Control with RStudio	

Analysis oooooooooooooooooo Writing 00000000

Using git in Rstudio

- Navigate to the Git tab and click Commit
- Check the boxes next to all *.R, and *.md files
- Write 'initial commit' in the box and click Commit



Analysis

Writing 00000000

Tracking changes with git

- Edit various files and commit the changes
- Click on the Git tab, then on the clock icon to view your commit history (project timeline)
- You can view the full project history, or review changes to a particular file
- You can continue working in this self-contained way (i.e., not putting anything online) and track the entire history of your project

Avoid tracking any large datasets or private info. These can be ignored by listing them by name in a .gitignore file

Analysis

Writing 00000000

Tracking changes with git

•) 😑 😑 RStudio: Review Chan	ges		
	Changes History master - (all commits) - 🕝		Q, Search	🛛 🛛 🖊 Pull
	Subject	Author	Date	SHA
Ŷ	(HEAD -> refs/heads/master) update comment for year conversion			
þ	add new section	Hansen Johnson <hansen.johnson@dal.ca></hansen.johnson@dal.ca>	2019-06-11	dccf276e
þ	change description	Hansen Johnson <hansen.johnson@dal.ca></hansen.johnson@dal.ca>	2019-06-11	96e03fe1
9	start using git!	Hansen Johnson <hansen.johnson@dal.ca></hansen.johnson@dal.ca>	2019-06-11	da5bc69b

	🚯 🚯 Commits 1-4 of 4 🕟 🛞 🕅
SHA	a446dacb
Author	Hansen Johnson <hansen.johnson@dal.ca></hansen.johnson@dal.ca>
Date	2019-06-11 07:59
Subject	update comment for year conversion
Parent	dccf276e
© src/	process_data.R
🤨 sro	/process_data.R View file @ a446dacb
	00 -17,7 +17,7 00 df = read.csv(infile, skip = 2, col.names = c("year", "ice_cover"))
17 17	# remove missing values
18 18	df = df[complete.cases(df),]
19 19	
20	# format year
20	# convert year to numeric
21 21	df\$year = as.numeric(as.character(df\$year))
22 22	
23 23	# save

Writing 00000000 Documentation

What is GitHub?

GitHub

- GitHub is not git
- GitHub is a massive hosting service for git repositories
- Provides convenient tools for reviewing and collaborating on code (and free backups!)
- Unlimited free public and private* repositories

* Only with \leq 3 collaborators (student accounts are unlimited)

Motivation	
0000	

Analysis

Writing 00000000 Documentation

Creating and linking with GitHub repository

- Go to GitHub user page
- Create a new repository with the same name as our example project (e.g., ice_cover)
- Ohoose to initialize without a readme

Owner	Repository name *	
🛃 hansenjohnson 🗸	/ ice_cover	
Great repository names ar	re short and memorable. Need inspiration? How about upgraded-spoon?	
Description (optional)		
Simple example project	to review some helpful tools / concepts for research	
Public Anyone can see this	repository. You choose who can commit.	
Public Anyone can see this Private You choose who can	repository. You choose who can commit. I see and commit to this repository.	
Public Anyone can see this Private You choose who can	repository. You choose who can commit. see and commit to this repository.	
Public Anyone can see this Private You choose who can Skip this step if you're imp Initialize this reposito	repository. You choose who can commit. 1 see and commit to this repository. porting an existing repository. Providth a README	
Public Anyone can see this Private You choose who can Skip this step if you're imp Initialize this reposito This will let you immediate	repository. You choose who can commit. 1 see and commit to this repository. porting an existing repository. by with a FEADME by clone the repository to your computer.	

Notivation

Writing 00000000

Creating and linking with GitHub repository

- Copy code listed in "... or push an existing repository from the command line"
- Ove to Rstudio and open Tools -> Terminal -> New Terminal
- Paste the lines into the terminal
- Sefresh your browser and check out your project online!





Analysis

Writing 00000000

Using GitHub

- Make commits on your computer
- When ready, push commits to GitHub by clicking on Push arrow on the git tab in Rstudio
- Oheck out new code online

La hansenjohnson / ice_cover									% Fork 0	
<> Code	Dissues 0 11 Pt	ull requests 0	Projects 0	🗉 Wiki	C Security	Insights 🔅 Se	ettings			
Simple example project to review some helpful tools / concepts for research Manage topics Edit										
⑦ 4 commits			β 1 branch		S 0 releas	S 0 releases		La 1 contributor		
Branch: master	New pull reque	est			Create new	file Upload files	Find File	Clone	or download 🗕	
Latest commit a446dac 13 minutes ago										
src	update comment for year conversion					13 minutes ago				
master.R		start using git!						17	' minutes ago	
i readme.me	d	add new sectio	n					14	minutes ago	



- Project contributors (collaborators, or you working on another computer) can clone the project onto their computer, commit changes, then push back to GitHub
- git and GitHub have many, many features for organization and collaboration including:
 - Branching
 - Merging / pull requests
 - Issue tracking
 - Website hosting

Check out fantastic GitHub documentation: https://guides.github.com

Analysis

Writing 00000000 Documentation

Key concepts

- Use readme files to describe your project, even if just to yourself
- Use git in Rstudio to track changes
- Use GitHub for backups, sharing, and collaboration

Possible next steps

- Dig deeper into git features (branching, pull requests, merging, etc)
- Use git and GitHub for collaboration
- Use Jekyll or Hugo to build project websites and host on GitHub

Analysis

Writing 00000000 Documentation

Questions?

Thanks to:

Christoph Renkl, Dalhousie Oceanography Student Association (DOSA), Methods in Ten Minutes (MTM), MEOPAR-WHaLE, and more!

Link to presentation:

https://hansenjohnson.org/talk/2019_meopar_atm/

Link to example project:

https://github.com/hansenjohnson/ice_cover_example/

Get in touch:

hansen.johnson@dal.ca