

Data Management Git & GitHub

ENTMLGY 6707
Entomological Techniques and Data Analysis

Look for efficient solutions



Learning objectives

Become familiar with best practices in data management

Compare and contrast approaches to data organization using spreadsheets

Given a data structure/input, anticipate barriers to loading the data into R

Introduction to Git/GitHub and its value for open science

Messy data

- Quite likely, in your previous math/stats course(s), you worked with data in homework problems provided by the instructor in a textbook
- The answers were in the back of the book
- It was tidy

Now, you will be analyzing “real life” data.

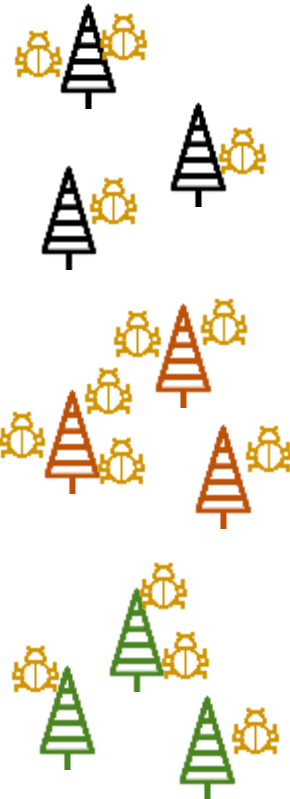
Messy data

- Quite likely, in your previous math/stats course(s), you worked with data in homework problems provided by the instructor in a textbook
- The answers were in the back of the book
- It was tidy

Now, you will be analyzing “real life” data.



Created by icongeek
from Noun Project



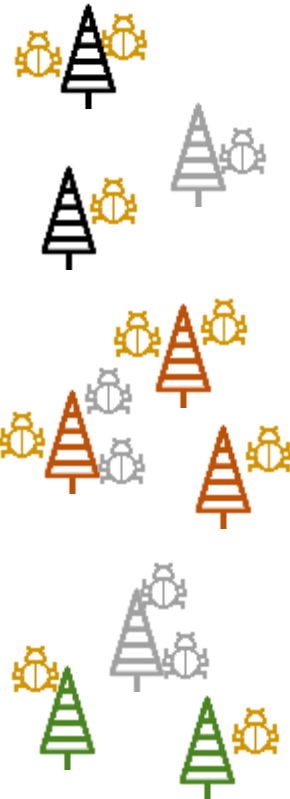
Messy data

- Quite likely, in your previous math/stats course(s), you worked with data in homework problems provided by the instructor in a textbook
- The answers were in the back of the book
- It was tidy

Now, you will be analyzing “real life” data.



Created by icongeek
from Noun Project



Best practices

Keep multiple copies of your data: hard and electronic

- Electronic = scanned hard copies & spreadsheet

MAKE SURE YOUR ELECTRONIC COPY IS BACKED UP AT ALL TIMES

Data safety issues are especially important when working with human subjects

- Understand custodial issues in data sharing in advance
- Ways to share sensitive data: removing personal information, create unique IDs

Spreadsheets

Typically, data is entered into a spreadsheet before exporting to a dedicated statistics program.

The most common spreadsheet is Microsoft Excel, but there are others (e.g., Google Docs, LibreOffice)

Proper data setup early in your investigation will avoid a lot of headaches in the future!

Data in spreadsheets

Avoid making “pretty” datasheets. Statistics programs, as a rule, don’t do “pretty” very well.

Each line of data contains all the variables **for a single observation**

Try to use a column for each variable

Don't do this

[illegible]

Bad example

Microsoft Excel								
File Home Insert Page Layout Formulas Data Review View Help Acrobat Search								
Paste	Cut Copy Format Painter	Calibri 11 A A	B I U Font Color Background Color	Alignment Merge & Center	General \$ % Number	Conditional Formatting	Format as Table	Normal Bad Check Cell Explanatory ...
N10	X ✓ fx							
	A	B	C	D	E	F	G	H
1	Info	State	County	Lon	Lat	Date	Survived? (1=yes, 0=no)	Weight in milligrams
2	Site1-Ind1-M	MS	Oktibbeha	-88.81	33.41	14-Feb	1	10
3	Site1-Ind2-Male	MS	Oktibbeha	-88.81	33.41	14-Feb	1	11
4	Site1-Ind3-M	MS	Oktibbeha	-88.81	33.41	17-Feb	0	10
5	Site2-Ind1-F	MS	Oktibbeha	-89.07	33.54	14-Feb		
6	Site2-Ind2-female	MS	Oktibbeha	-89.07	33.54	17-Feb	0	8
7	Site3-Ind1-Female	MS	Choctaw	-89.25	33.41	14-Feb	0	9
8	Site4-Ind1-M	AL	Pickens	-88.25	33.31	14-Feb		12.5
9	Site4-Ind2-M	AL	Pickens	-88.25	33.31	17-Feb	1	11.5
10	Site5-Ind1-M	AL	Lamar	-88.09	33.71	17-Feb	1	11
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								

Good example

[illegible]

Codebook / Metadata

Data is typically kept along with a codebook or metadata. This is (more or less) information describing a particular study or data

The typical metadata contains:

1. A description of how the data were collected including sampling design
2. The variables contained in the data
3. In the case of surveys, the survey instrument or questionnaire used to solicit responses from the respondent and the coded values of each question
4. The format and/or units of each variable within the raw data file
5. Meaning of the coded values for each variable, including (as necessary) whether data are continuous, categorical, ordinal, nominal, binary, etc.

Quality control checks

Once your data is entered, it is a good idea to perform a quality control check before reading the file into a statistics program.

This is essential whether the data are generated by machine or people

Likely required to publish your data with manuscript

<http://www.ecography.org/authors/author-guidelines>



ECOGRAPHY
A JOURNAL OF SPACE AND TIME IN ECOLOGY
PUBLISHED BY THE NORDIC SOCIETY OIKOS.

AUTHOR GUIDELINES

This page explains how to prepare your manuscript for submission to the journal Ecography, a Nordic Society Oikos publication. Before submitting, please make sure that your article fits within the journal's [aims and scope](#).

Please prepare your manuscript carefully, following the guidelines on this page.

Data archiving statement

For articles published in Ecography, it is required that authors deposit data supporting their accepted papers in public archives of their choice (see section on **Data sharing and repositories** below). Authors must confirm that they deposit their data in a public repository and indicate the repository of their choice.

Some journals now require code, too

<https://www.esa.org/publications/data-policy/>

OVERVIEW

ESA has adopted a society-wide Open Research Policy for its publications to further support scientific exploration and preservation, allow a full assessment of published research, and streamline policies across our family of journals. An open research policy provides full transparency for scientific data and code, facilitates replication and synthesis, and aligns ESA journals with current standards. As of 1 February 2021, all new manuscript submissions to ESA journals must abide by the following policy.

As a condition for publication in ESA journals, all underlying data and novel statistical code pertinent to the results presented in the publication must be made available in a permanent, publicly accessible data archive or repository upon acceptance of a manuscript, with rare exceptions (see the “Details” tab for more information). Archived data and novel statistical code should be sufficiently complete to allow replication of tables, graphs, and statistical analyses reported in the original publication, and perform new or meta-analyses. As such, the desire of authors to control additional research with these data and/or code shall not be grounds for withholding material.



Git and GitHub – What are they?

Git: Software that handles version control on your repository

- Working in the background when using GitHub

GitHub: Web interface that hosts your repository online

- Allows for collaboration on projects
- Interfaces with R/RStudio & Git





Kayla I Perry

kiperry

Edit profile

3 followers · 10 following

kiperry1488@gmail.com

<https://u.osu.edu/perrylab/>

https://www.researchgate.net/profile/Kayla_Perry

Organizations



Popular repositories

Customize your pins

[ENT6702_DataAnalysis](#)

Public

Repository for the course ENT 6702: Entomological Techniques and Data Analysis

R 1

[Ash_Beetle_Communities](#)

Public

R

[Arthropod_Marking](#)

Public

[LadyBeetle_SEM](#)

Public

R

[ESA_NCB_2021](#)

Public

R

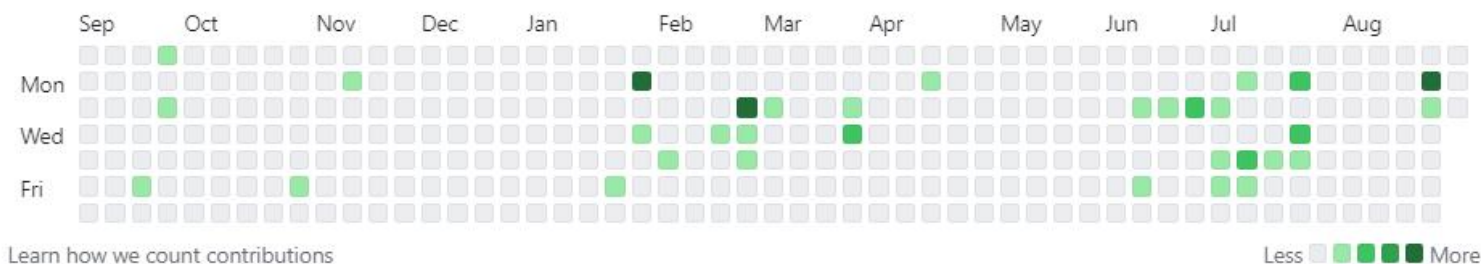
[OSU_Statistics_Workshops](#)

Public

1

72 contributions in the last year

Contribution settings



Contribution activity

2023

August 2023

2022

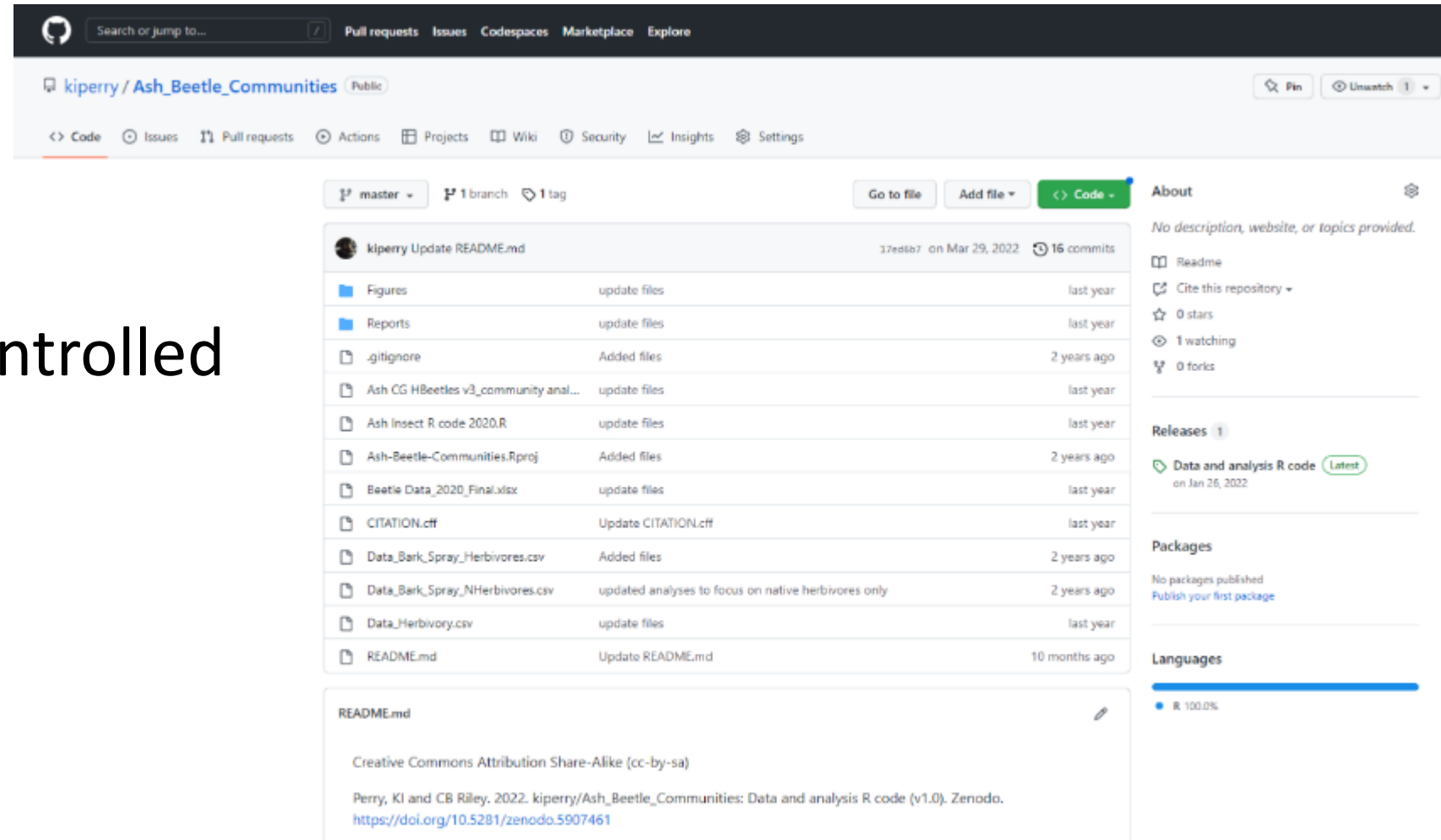
What is a repository (or repo)?

Place for all files associated with a project

With GitHub, your repo lives on your computer and online

Each file is version controlled with documented development history

Public or private





Kayla I Perry

kiperry

Edit profile

3 followers · 10 following

kiperry1488@gmail.com

<https://u.osu.edu/perrylab/>

https://www.researchgate.net/profile/Kayla_Perry

Organizations



Find a repository...

Type ▾

Language ▾

Sort ▾

New

ENT6702_DataAnalysis Public

Repository for the course ENT 6702: Entomological Techniques and Data Analysis

R 1 Updated 4 days ago

Star ▾

OSU_Statistics_Workshops Public

1 Updated last week

Star ▾

Ohio_Bees Private

HTML Updated on Jul 27

Star ▾

IDH_Ground_Beetles Public

R Updated on Jul 20

Star ▾

WI_Bumble_Bees Private

R Updated on Jun 20

Star ▾

PNR_Beetles Public

Star ▾

Make changes or updates to repo with commit

Save a version of a file, and provide notes on what you changed

When you commit a file in Git/GitHub, you are saving a new version, but also keeping a record of the changes you made



Commit changes

Create README.md

Add an optional extended description...

kiperry1488@gmail.com

Choose which email address to associate with this commit

☒ Commit directly to the `master` branch.

☐ Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Commit changes

Cancel

Commits on Jan 26, 2022		
Update CITATION.cff	kiperry committed on Jan 26, 2022	Verified 9e13ee0
Update README.md	kiperry committed on Jan 26, 2022	Verified b0b11db
Create CITATION.cff	kiperry committed on Jan 26, 2022	Verified 8ecf12d
Update README.md	kiperry committed on Jan 26, 2022	Verified 35164c5
Merge branch 'master' of https://github.com/kiperry/Ash-Beetle-Commun...	kiperry committed on Jan 26, 2022	91a335c
update files	kiperry committed on Jan 26, 2022	329f81d
Commits on Nov 29, 2021		
Update README.md	kiperry committed on Nov 29, 2021	Verified 2c6297e
Commits on Jun 15, 2020		
updated analyses to focus on native herbivores only	kiperry committed on Jun 15, 2020	2bd4f4c
Commits on Jun 11, 2020		
Merge branch 'master' of https://github.com/kiperry/Ash-Beetle-Commun...	kiperry committed on Jun 11, 2020	853089b
Added files	kiperry committed on Jun 11, 2020	457acab
Delete Data_Bark_Spray_Herbivores.csv	kiperry committed on Jun 11, 2020	Verified 47c5b1d
Delete Beetle Data_2020_Final.xlsx	kiperry committed on Jun 11, 2020	Verified 75250e9

Make changes or updates to repo with commit

Save a version of a file, and provide notes on what you changed

When you commit a file in Git/GitHub, you are saving a new version, but also keeping a record of the changes you made



Commit changes

Create README.md

Add an optional extended description...

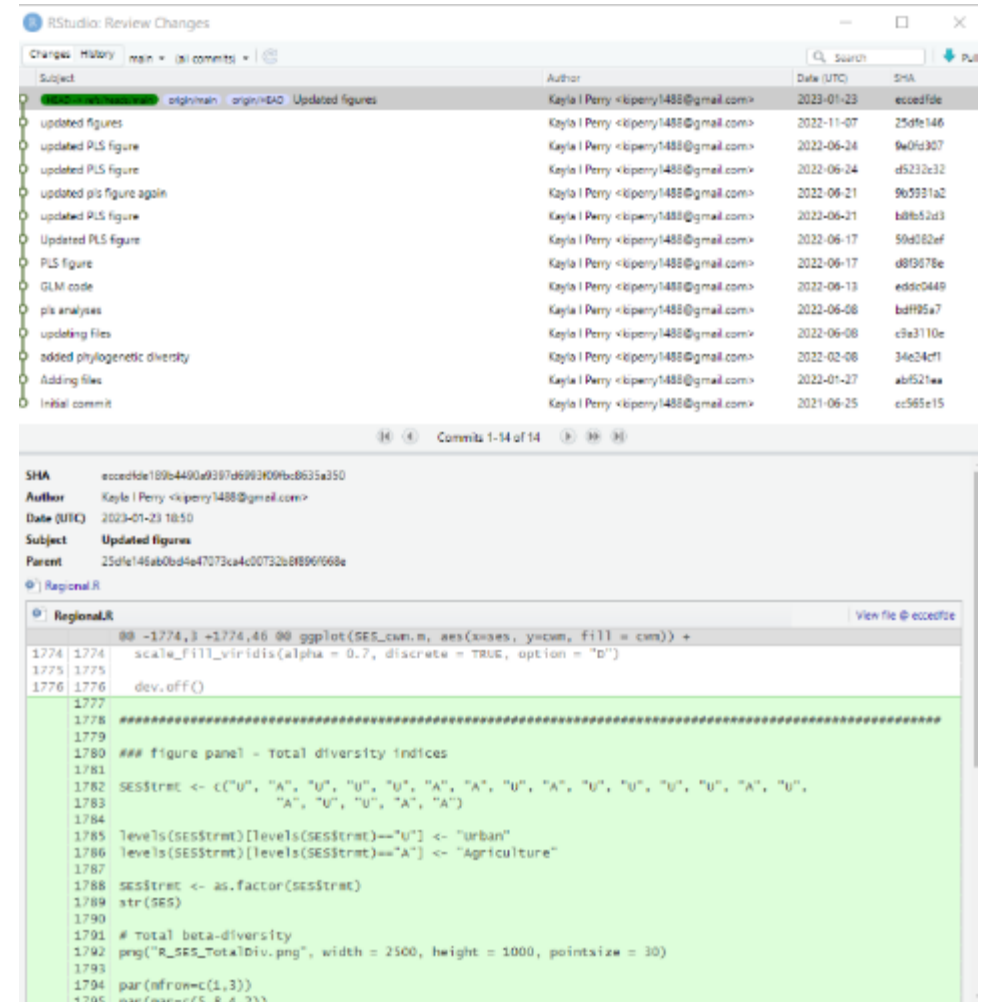
kiperry1488@gmail.com

Choose which email address to associate with this commit

☒ Commit directly to the master branch.

☐ Create a new branch for this commit and start a pull request. [Learn more about pull requests.](#)

Commit changes Cancel



RStudio Review Changes

Changes	History	main	(all commits)	Search	Pull
Updated figures	origin/main	origin/HEAD	Updated figures		
updated figures					
updated PLS figure					
updated PLS figure					
updated pls figure again					
updated PLS figure					
Updated PLS figure					
PLS figure					
GLM code					
pls analyses					
updating files					
added phylogenetic diversity					
Adding files					
Initial commit					

Commit 1-14 of 14

SHA eeced1e189b4490a9397d699200bc8635a350

Author Kayla I Perry <kiperry1488@gmail.com>

Date (UTC) 2023-01-23 18:50

Subject Updated figures

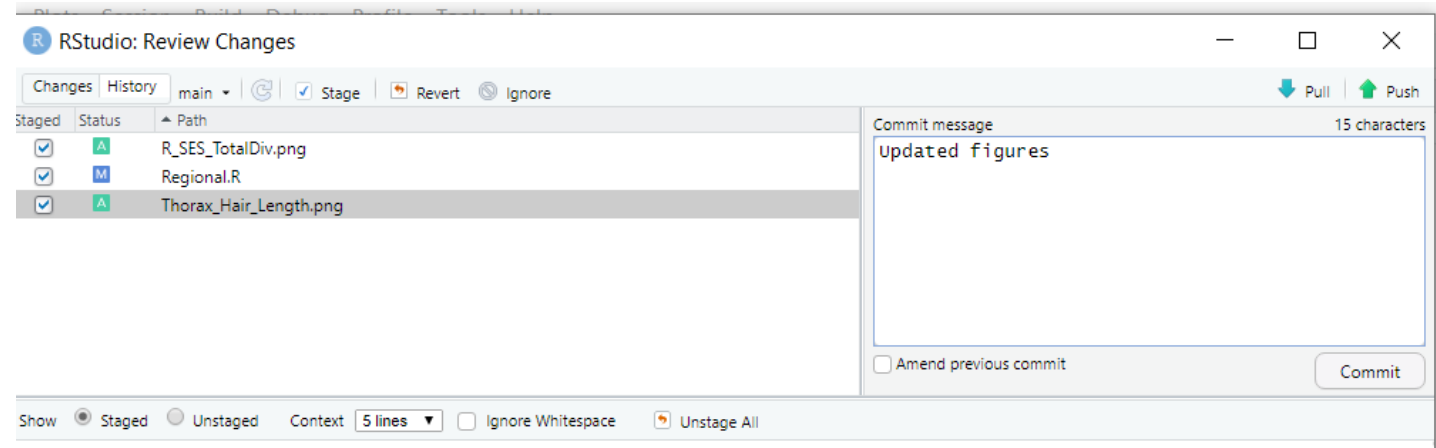
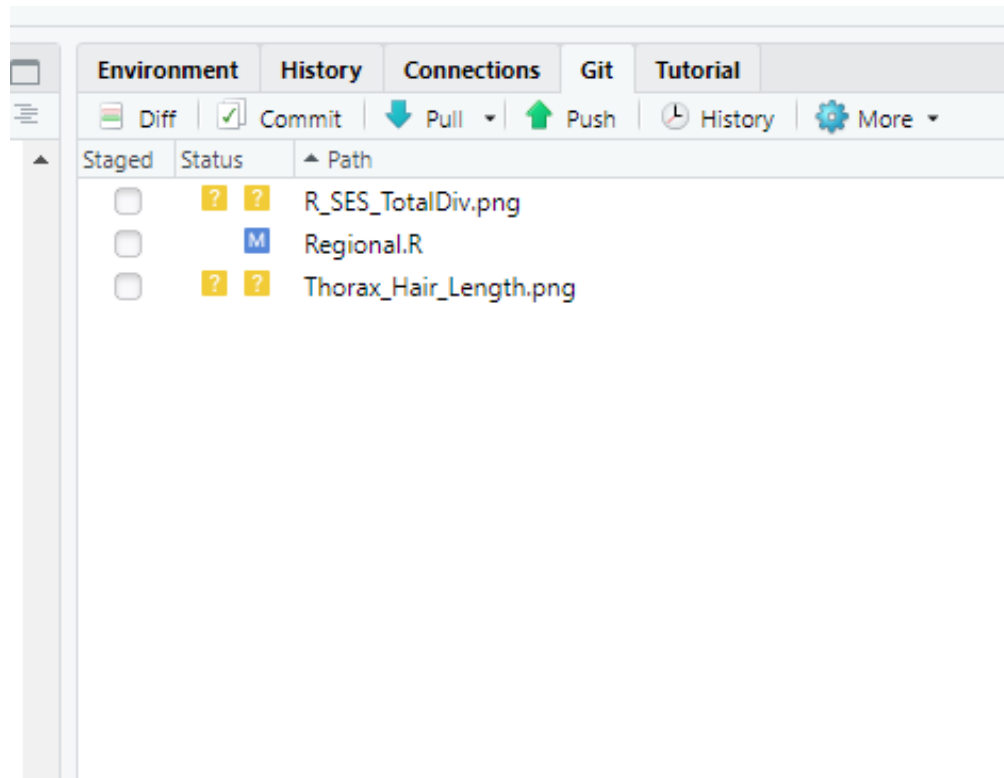
Parent 25dfe146ab0bd4e47073ca4c00732b68896f668e

Regional.R

```
## -1774,3 +1774,46 ## ggplot(SE$_cwm.m, aes(x=cwm, y=cwm, fill = cwm)) +
1774 1774 scale_fill_viridis(alpha = 0.7, discrete = TRUE, option = "d")
1775 1775
1776 1776 dev.off()
1777
1778 #####
1779
1780 ## figure panel - Total diversity indices
1781
1782 SES$trmt <- c("U", "A", "U", "U", "U", "A", "A", "U", "A", "U", "U", "U", "A", "U",
1783             "A", "U", "U", "A", "A")
1784
1785 levels(SES$trmt)[levels(SES$trmt)=="U"] <- "urban"
1786 levels(SES$trmt)[levels(SES$trmt)=="A"] <- "Agriculture"
1787
1788 SES$trmt <- as.factor(SES$trmt)
1789 str(SES)
1790
1791 # Total beta-diversity
1792 png("R_SES_TotalDiv.png", width = 2500, height = 1000, pointsize = 30)
1793
1794 par(mfrow=c(1,3))
1795 par(mar=c(5,8,4,2))
```

Pull, Commit, then Push

1. **Pull** from the online repository to update your local files
2. **Commit** to save a new version of a file(s)
3. **Push** those changes online to the repository



1) Sync project files locally on your computer and online

2) Make commits to record changes to files over time

3) Facilitates remote collaboration within a project repository



What can we do with Git/GitHub?

- 1) Experiment on projects without breaking them – **Branch**
- 2) Make, assign, and keep track of tasks – **Issues**
- 3) Access existing projects made by others – **Fork or Clone**
- 4) Build on existing projects with collaborators – **Pull, Commit, Push**

Why use Git/GitHub with R?

Facilitates research transparency and reproducibility

Share data, code, and analyses with collaborators and scientific community

- Track development history over time

Aligns with open science journal requirements



Submitting your manuscript for review?

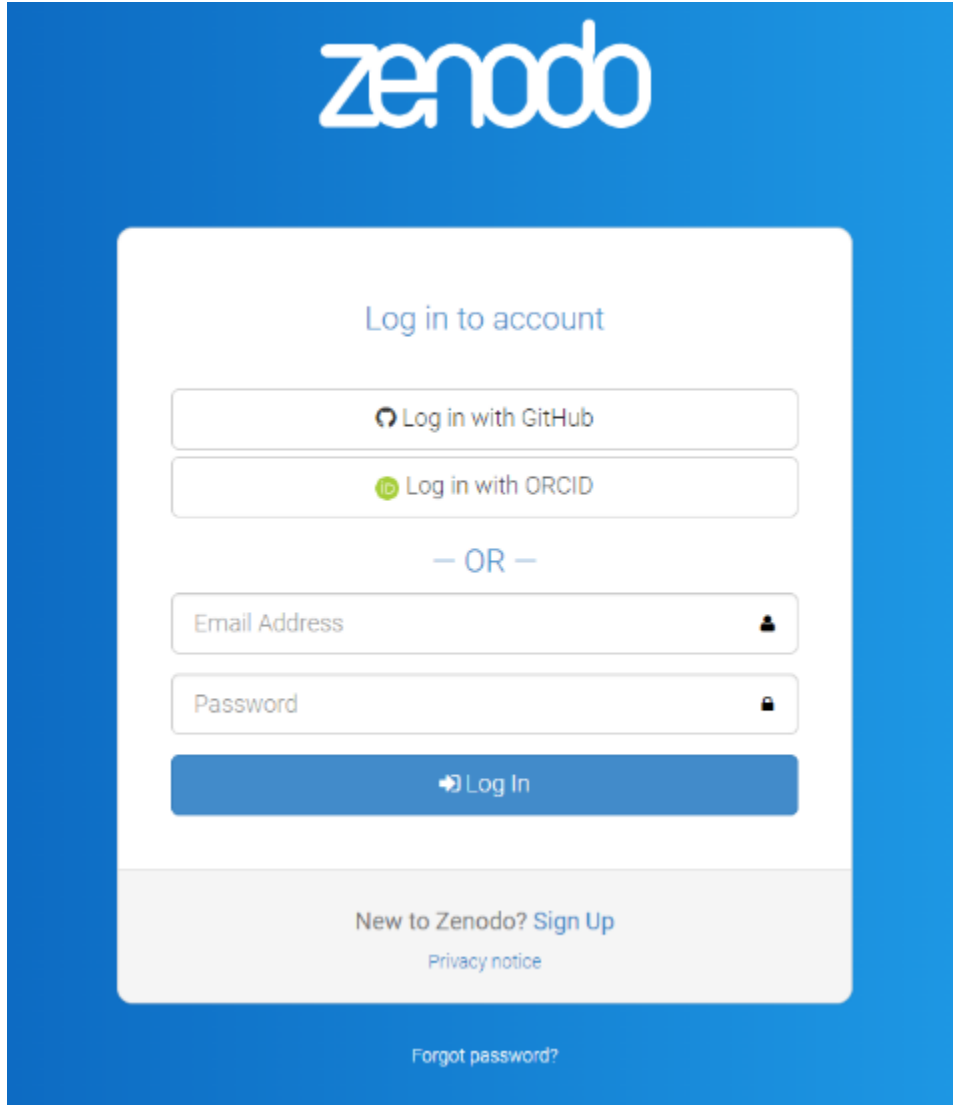
Include your GitHub repository link in your open research/data availability statement

Many journals now require submission of data and code for peer-review

23 **Open Research Statement:** Data are already published and publicly available, with those items properly
24 cited in this submission. This submission uses novel code, which is provided in an external repository to
25 be evaluated during the peer review process and are available at
26 https://github.com/BahlaiLab/Ohio_ladybeetles. If this paper is accepted for publication, data and code
27 will be permanently archived in Zenodo.

Link GitHub repository with Zenodo for DOI

Developed under European OpenAIRE Program
Operated by CERN



The image shows the Zenodo login page. At the top is the Zenodo logo. Below it is a 'Log in to account' section. There are two buttons: 'Log in with GitHub' and 'Log in with ORCID'. Below these is a separator '— OR —'. Then there are input fields for 'Email Address' and 'Password', each with an icon (a person for email, a lock for password). Below the inputs is a blue 'Log In' button. At the bottom, there is a link for 'New to Zenodo? Sign Up' and a link for 'Privacy notice'. At the very bottom is a link for 'Forgot password?'.

zenodo

Log in to account

Log in with GitHub

Log in with ORCID

— OR —

Email Address

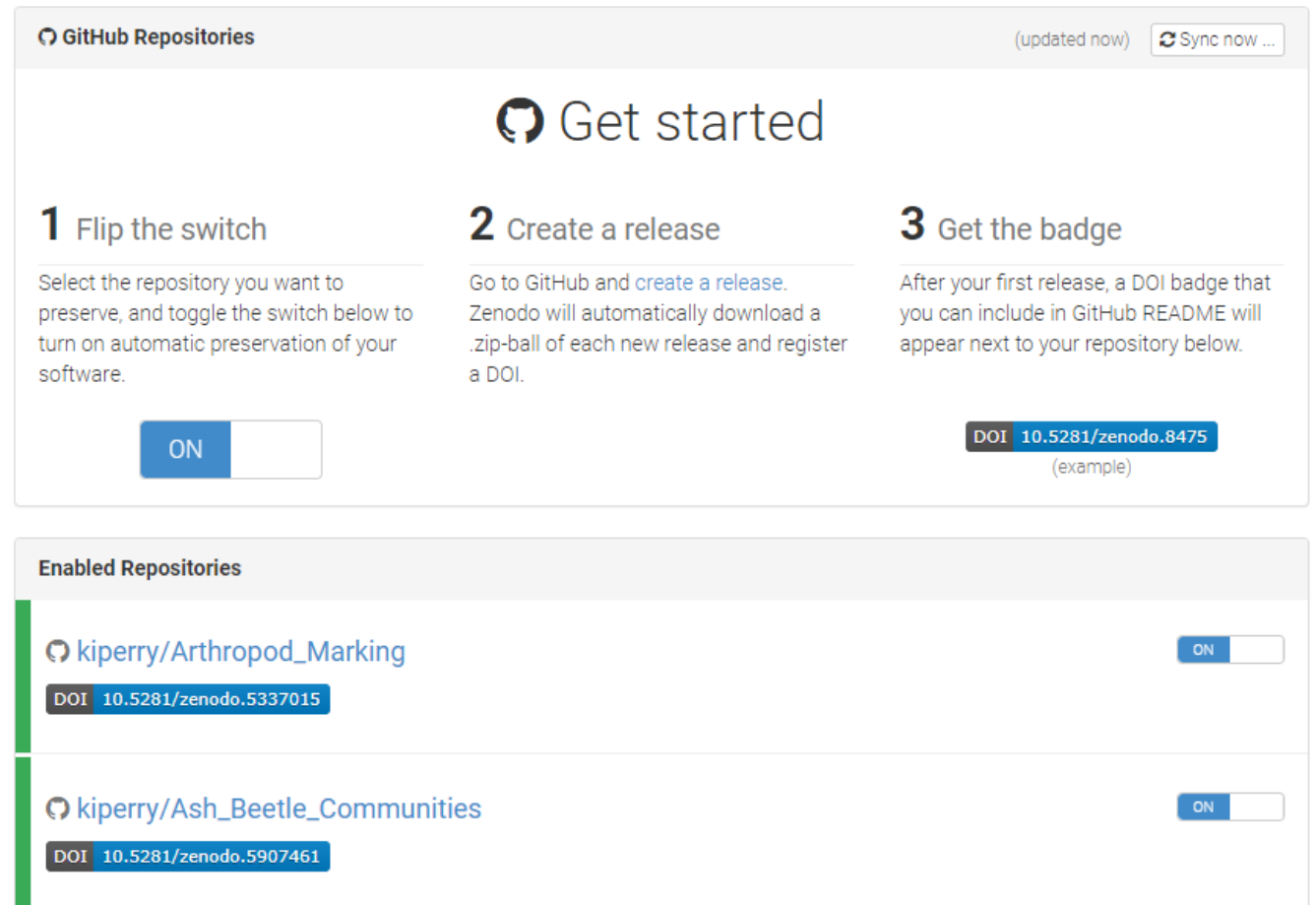
Password

Log In

New to Zenodo? Sign Up

Privacy notice

Forgot password?



The image shows the 'GitHub Repositories' management page in Zenodo. At the top, it says 'GitHub Repositories' with a '(updated now)' status and a 'Sync now ...' button. Below this is a 'Get started' section with three steps: 1. Flip the switch, 2. Create a release, and 3. Get the badge. Step 1 includes a toggle switch labeled 'ON'. Step 2 includes instructions to go to GitHub and create a release. Step 3 includes instructions to get a DOI badge and an example DOI: 10.5281/zenodo.8475. Below the 'Get started' section is an 'Enabled Repositories' section. It lists two repositories: 'kiperry/Arthropod_Marking' and 'kiperry/Ash_Beetle_Communities'. Each repository has a toggle switch labeled 'ON' and a DOI badge: 10.5281/zenodo.5337015 and 10.5281/zenodo.5907461 respectively.

GitHub Repositories (updated now) Sync now ...

Get started

1 Flip the switch

Select the repository you want to preserve, and toggle the switch below to turn on automatic preservation of your software.

ON

2 Create a release

Go to GitHub and [create a release](#). Zenodo will automatically download a .zip-ball of each new release and register a DOI.

3 Get the badge

After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.

DOI 10.5281/zenodo.8475 (example)

Enabled Repositories

kiperry/Arthropod_Marking ON DOI 10.5281/zenodo.5337015

kiperry/Ash_Beetle_Communities ON DOI 10.5281/zenodo.5907461

Software and Data Products category on your CV!

January 26, 2022

Dataset

Open Access

Edit

New version

44

views

2

downloads

[See more details...](#)

Study data and analysis code

 Kayla I Perry;  Christopher B Riley

Study data and R code, first release v1.0

This dataset supports the following study:

Perry, KI, CB Riley, F Fan, J Radl, DA Herms, and MM Gardiner. The value of hybrid and nonnative ash for the conservation of ash specialists is limited following late stages of emerald ash borer invasion, Agricultural and Forest Entomology, doi.org/10.1111/afe.12499

Creative Commons Attribution Share-Alike (cc-by-sa)




Preview

Ash_Beetle_Communities-v1.0.zip

kiperry-Ash_Beetle_Communities-91a335c

◦ .gitignore	40 Bytes
◦ Ash CG HBeetles v3_community analyses.R	35.2 kB
◦ Ash Insect R code 2020.R	6.8 kB
◦ Ash-Beetle-Communities.Rproj	209 Bytes
◦ Beetle Data_2020_Final.xlsx	91.5 kB
◦ Data_Bark_Spray_Herbivores.csv	25.1 kB
◦ Data_Bark_Spray_NHerbivores.csv	23.1 kB
◦ Data_Herbivory.csv	27.8 kB
◦ Figures	
▪ NBetaDiversity_Bark_Ash.png	10.1 kB
▪ NBetaDiversity_Spray_Ash.png	10.3 kB
▪ NMDS_Herbivores_Bark_Ash.png	34.3 kB
▪ NMDS_Herbivores_Spray_Ash.png	32.9 kB
▪ NMDS_NHerbivores_Canopy_Bark_Ash.png	64.1 kB
▪ NSpecies_Rarefaction_Ash.png	27.4 kB
◦ README.md	257 Bytes
◦ Reports	

Files (428.0 kB)

Name	Size	
kiperry/Ash_Beetle_Communities-v1.0.zip	428.0 kB	 Preview  Download
md5:55e0cc39fb6c5ee28ef1a4ffd5c4414e 		

Available in

GitHub

Indexed in

OpenAIRE

Publication date:

January 26, 2022

DOI:

DOI 10.5281/zenodo.5907461

Published in:

Agricultural and Forest Entomology;
doi.org/10.1111/afe.12499.

Related identifiers:

Supplement to
https://github.com/kiperry/Ash_Beetle_Communities/tree/v1.0

License (for files):

 Creative Commons Attribution Share Alike 4.0 International














Update readme file on GitHub with DOIs

kiperry / Ash_Beetle_Communities Public


[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

master 1 branch 1 tag

[Go to file](#) [Add file](#) [Code](#)

	kiperry Update README.md	17ed6b7 on Mar 29, 2022	16 commits
	Figures	update files	last year
	Reports	update files	last year
	.gitignore	Added files	2 years ago
	Ash CG HBeetles v3_community anal...	update files	last year
	Ash Insect R code 2020.R	update files	last year
	Ash-Beetle-Communities.Rproj	Added files	2 years ago
	Beetle Data_2020_Final.xlsx	update files	last year
	CITATION.cff	Update CITATION.cff	last year
	Data_Bark_Spray_Herbivores.csv	Added files	2 years ago
	Data_Bark_Spray_NHerbivores.csv	updated analyses to focus on native herbivores only	2 years ago
	Data_Herbivory.csv	update files	last year
	README.md	Update README.md	10 months ago

README.md




Creative Commons Attribution Share-Alike (cc-by-sa)


Perry, KI and CB Riley. 2022. kiperry/Ash_Beetle_Communities: Data and analysis R code (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5907461>


Perry, KI, CB Riley, F Fan, J Radl, DA Herms, and MM Gardiner. The value of hybrid and nonnative ash for the conservation of ash specialists is limited following late stages of emerald ash borer invasion, Agricultural and Forest Entomology, <https://doi.org/10.1111/afe.12499>


About


No description, website, or topics provided.

 Readme


 Cite this repository

 0 stars

 1 watching

 0 forks

Releases 1

 Data and analysis R code Latest
on Jan 26, 2022

Packages

No packages published
[Publish your first package](#)

Languages

R 100.0%

Submitting your manuscript for review?

Include your GitHub repository link in your open research/data availability statement

23 **Open Research Statement:** Data are already published and publicly available, with those items properly
24 cited in this submission. This submission uses novel code, which is provided in an external repository to
25 be evaluated during the peer review process and are available at
26 https://github.com/BahlaiLab/Ohio_ladybeetles. If this paper is accepted for publication, data and code
27 will be permanently archived in Zenodo.

28 **Open Research Statement:** Ohio lady beetle records were compiled from 25 institutions, which are
29 listed in Appendix S1: Table S1. Query details for obtaining these data are provided in Appendix S1:
30 Section S1. Data and code that support our findings are archived in Zenodo,
31 <https://doi.org/10.5281/zenodo.11263088>

R Assignment – Loading data into R

- When you have a dataset ready for analysis, it needs to be exported to a statistics program
- Some, but not all programs, can read an Excel file directly
- More commonly, the data needs to be saved as a plain text file (space or tab delimited)
- **Understand how the file is delimited and what the statistics program is expecting!**



Files and delimitations:

- *.txt Plain text
- *.txt Tab-delimited plain text
- *.dat Space delimited plain text
- *.csv Comma separated values