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

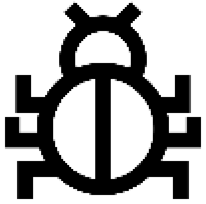
- Quite likely, in your previous math/stats course(s), you worked with data in homework problems provided by the instructor in a textbook
- The answers were in the back of the book
- It was tidy

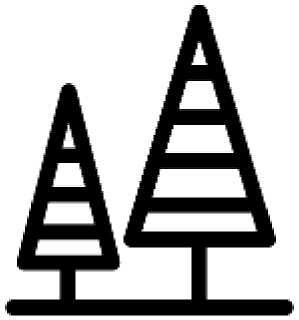
Now, you will be analyzing “real life” data.

Messy data

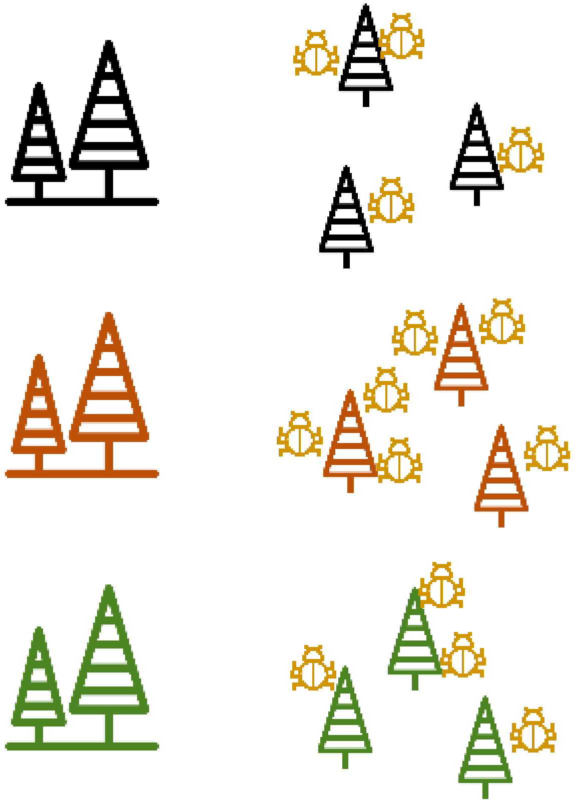
- Quite likely, in your previous math/stats course(s), you worked with data in homework problems provided by the instructor in a textbook
- The answers were in the back of the book
- It was tidy

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





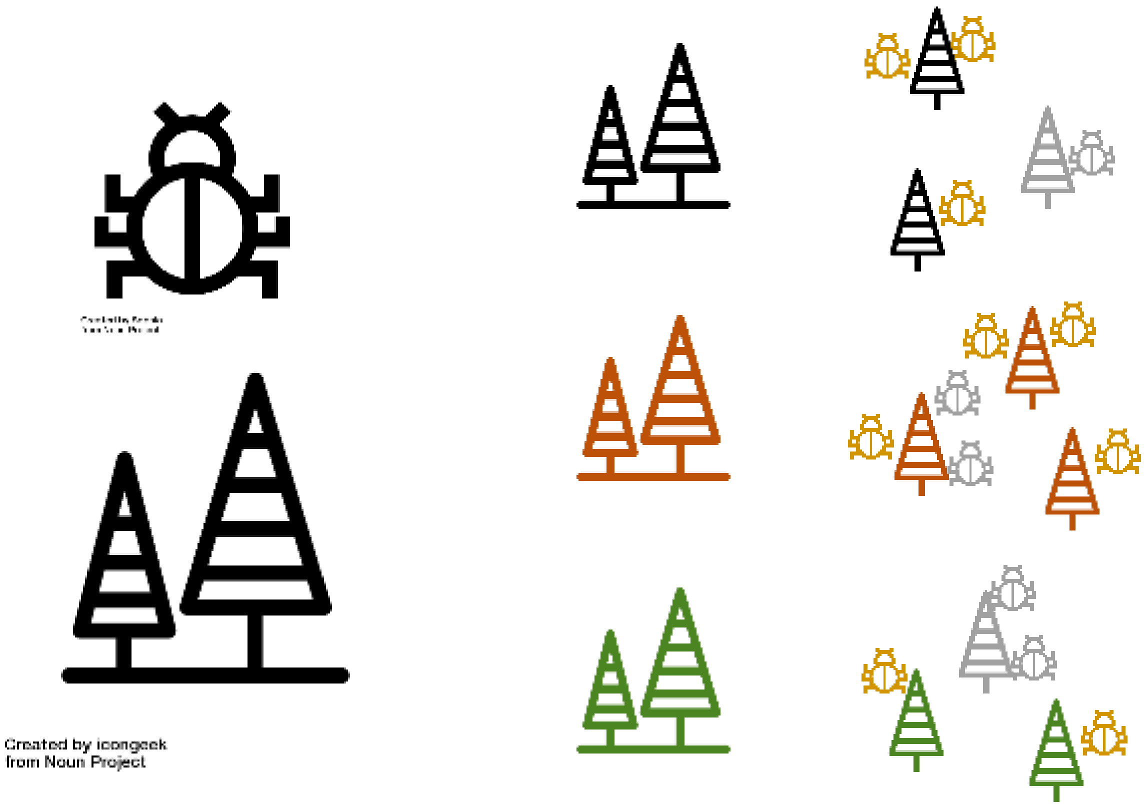
Created by icongeek
from Noun Project



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 [illegible] from Noun Project

Created by icongeek from Noun Project

Best practices

Keep multiple copies of your data: hard and electronic

- Electronic = scanned hard copies & spreadsheet

<u>MAKE SURE YOUR ELECTRONIC COPY IS BACKED UP AT ALL TIMES</u>

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 importing in 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!

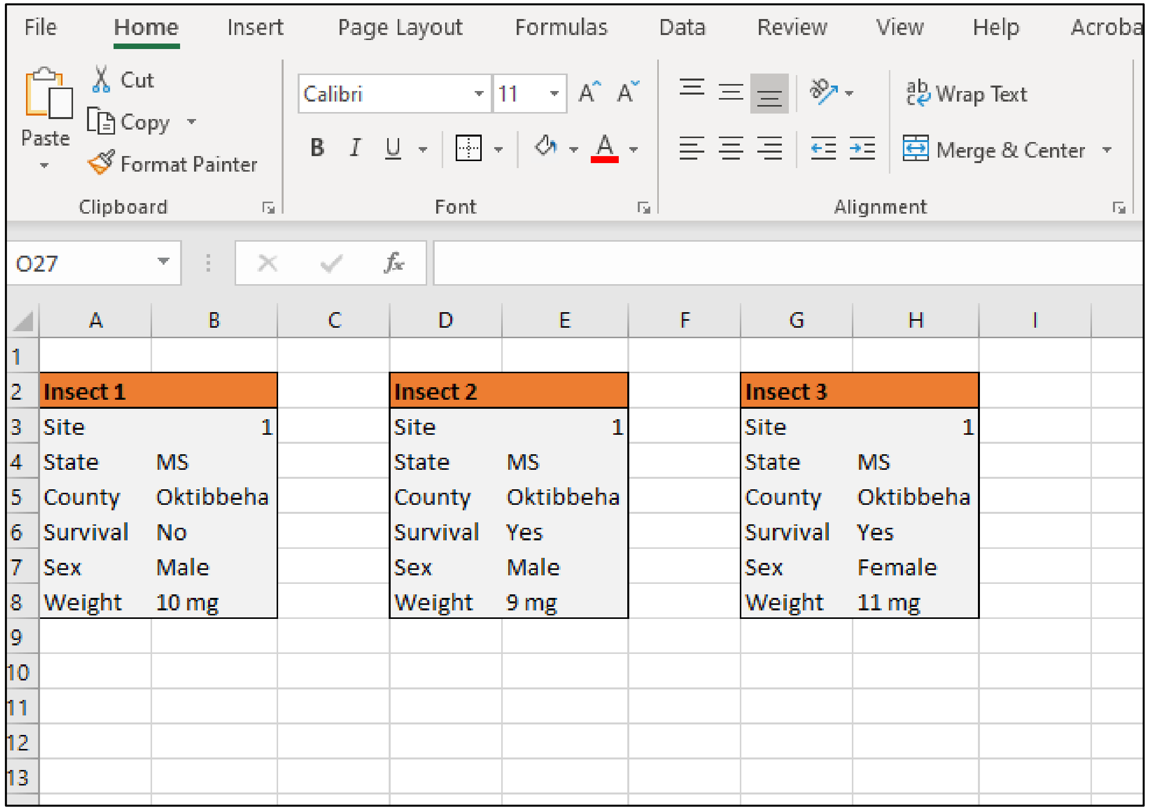
Data in spreadsheets

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**

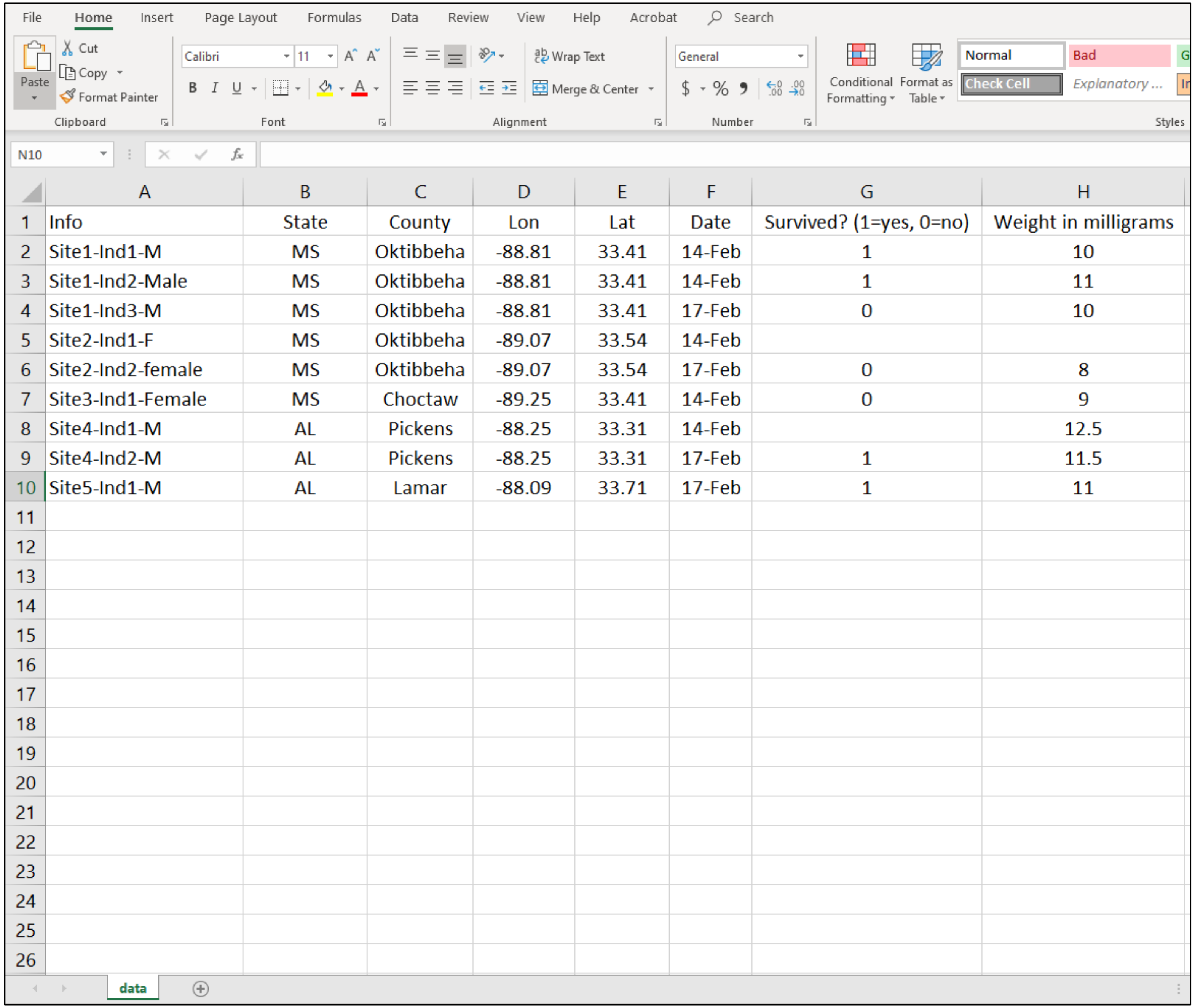
Try to use a column for each variable

Don’t do this



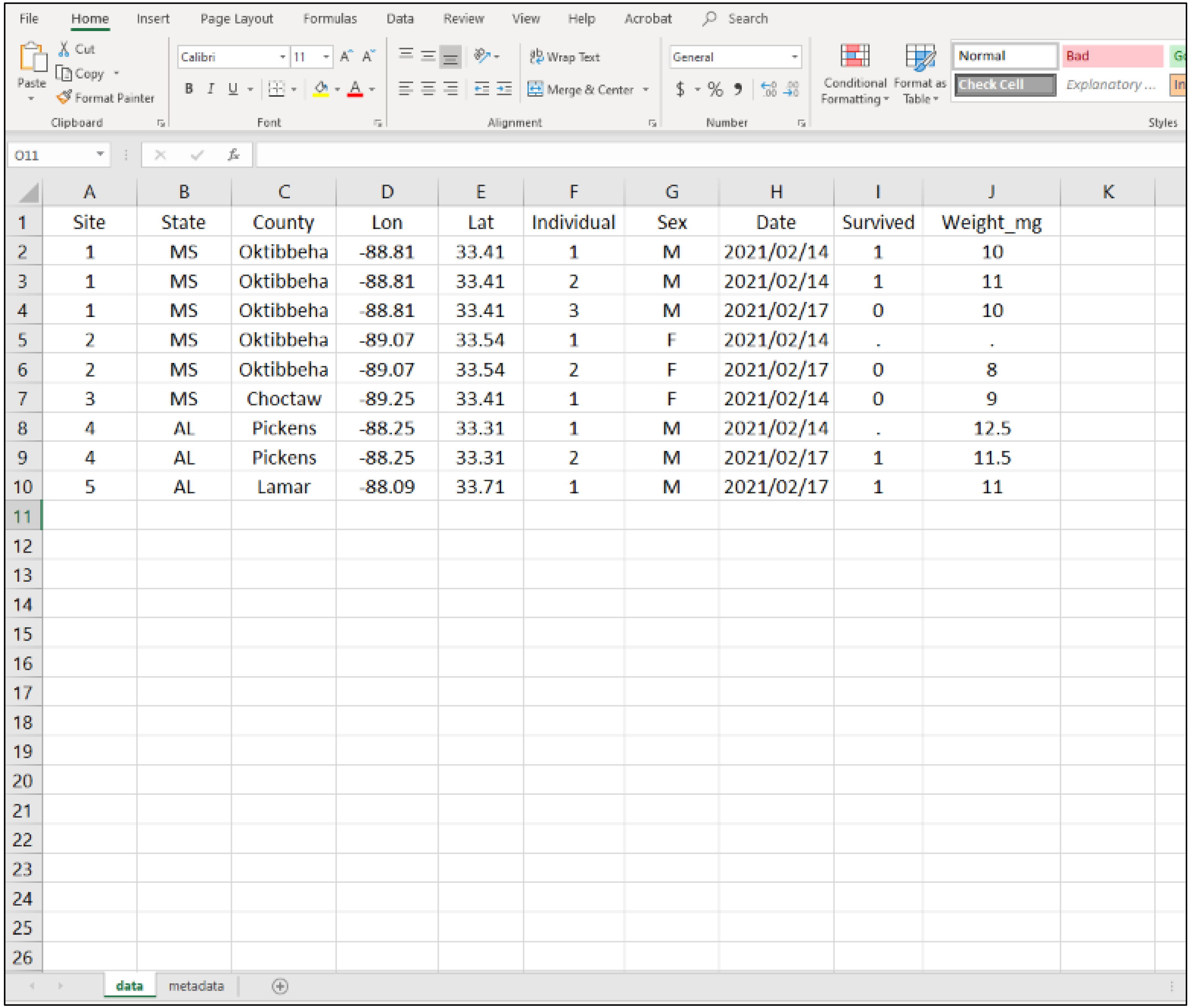
	A	B	C	D	E	F	G	H	I
1									
2	Insect 1			Insect 2			Insect 3		
3	Site	1		Site	1		Site	1	
4	State	MS		State	MS		State	MS	
5	County	Oktibbeha		County	Oktibbeha		County	Oktibbeha	
6	Survival	No		Survival	Yes		Survival	Yes	
7	Sex	Male		Sex	Male		Sex	Female	
8	Weight	10 mg		Weight	9 mg		Weight	11 mg	
9									
10									
11									
12									
13									

Bad example



Info	State	County	Lon	Lat	Date	Survived? (1=yes, 0=no)	Weight in milligrams
Site1-Ind1-M	MS	Oktibbeha	-88.81	33.41	14-Feb	1	10
Site1-Ind2-Male	MS	Oktibbeha	-88.81	33.41	14-Feb	1	11
Site1-Ind3-M	MS	Oktibbeha	-88.81	33.41	17-Feb	0	10
Site2-Ind1-F	MS	Oktibbeha	-89.07	33.54	14-Feb		
Site2-Ind2-female	MS	Oktibbeha	-89.07	33.54	17-Feb	0	8
Site3-Ind1-Female	MS	Choctaw	-89.25	33.41	14-Feb	0	9
Site4-Ind1-M	AL	Pickens	-88.25	33.31	14-Feb		12.5
Site4-Ind2-M	AL	Pickens	-88.25	33.31	17-Feb	1	11.5
Site5-Ind1-M	AL	Lamar	-88.09	33.71	17-Feb	1	11

Good example



Site	State	County	Lon	Lat	Individual	Sex	Date	Survived	Weight_mg
1	MS	Oktibbeha	-88.81	33.41	1	M	2021/02/14	1	10
1	MS	Oktibbeha	-88.81	33.41	2	M	2021/02/14	1	11
1	MS	Oktibbeha	-88.81	33.41	3	M	2021/02/17	0	10
2	MS	Oktibbeha	-89.07	33.54	1	F	2021/02/14	.	.
2	MS	Oktibbeha	-89.07	33.54	2	F	2021/02/17	0	8
3	MS	Choctaw	-89.25	33.41	1	F	2021/02/14	0	9
4	AL	Pickens	-88.25	33.31	1	M	2021/02/14	.	12.5
4	AL	Pickens	-88.25	33.31	2	M	2021/02/17	1	11.5
5	AL	Lamar	-88.09	33.71	1	M	2021/02/17	1	11

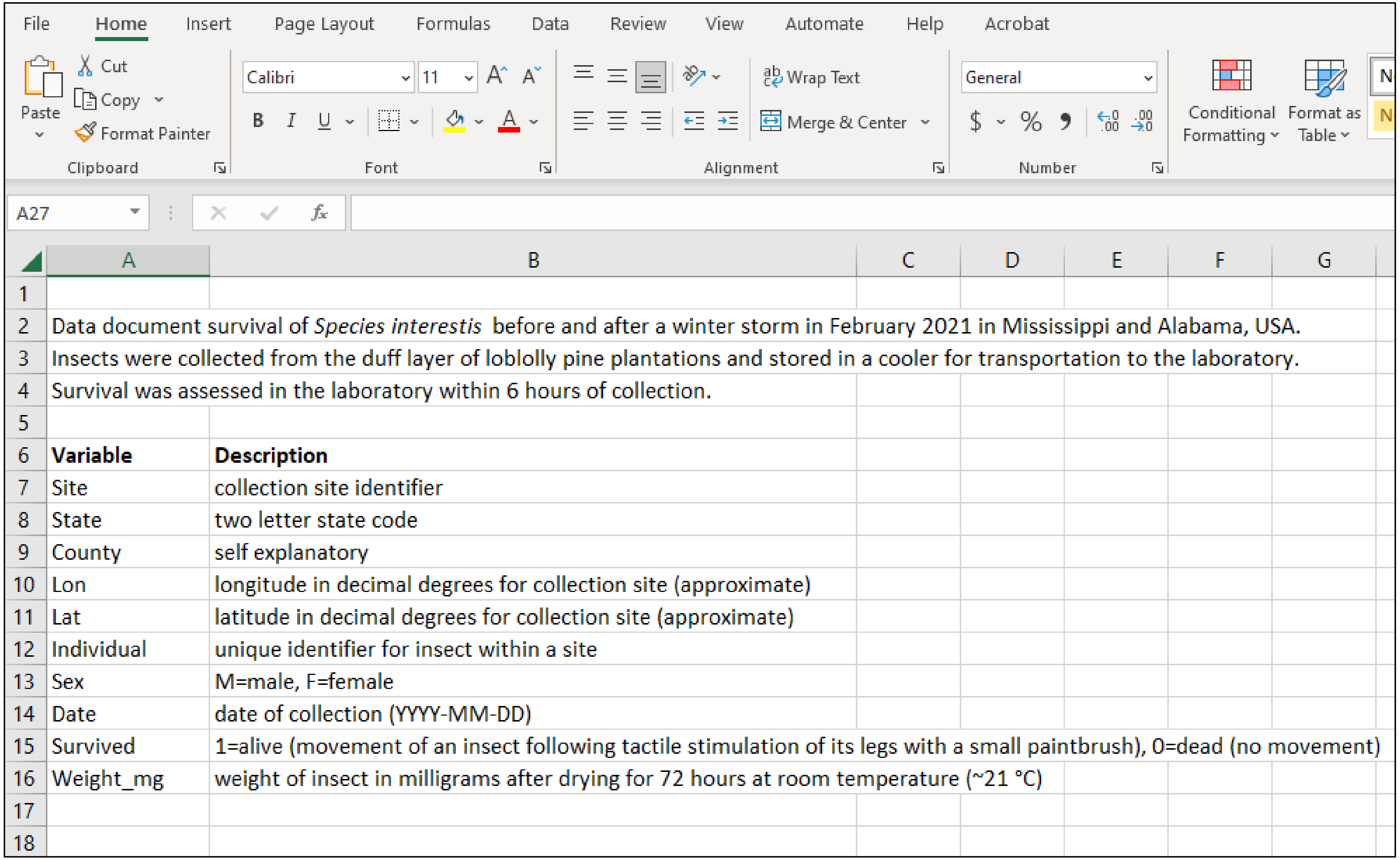
Codebook / Metadata

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.

Metadata



Data document survival of *Species interestis* before and after a winter storm in February 2021 in Mississippi and Alabama, USA.

Insects were collected from the duff layer of loblolly pine plantations and stored in a cooler for transportation to the laboratory.

Survival was assessed in the laboratory within 6 hours of collection.

Variable	**Description**
Site	collection site identifier
State	two letter state code
County	self explanatory
Lon	longitude in decimal degrees for collection site (approximate)
Lat	latitude in decimal degrees for collection site (approximate)
Individual	unique identifier for insect within a site
Sex	M=male, F=female
Date	date of collection (YYYY-MM-DD)
Survived	1=alive (movement of an insect following tactile stimulation of its legs with a small paintbrush), 0=dead (no movement)
Weight_mg	weight of insect in milligrams after drying for 72 hours at room temperature (~21 °C)

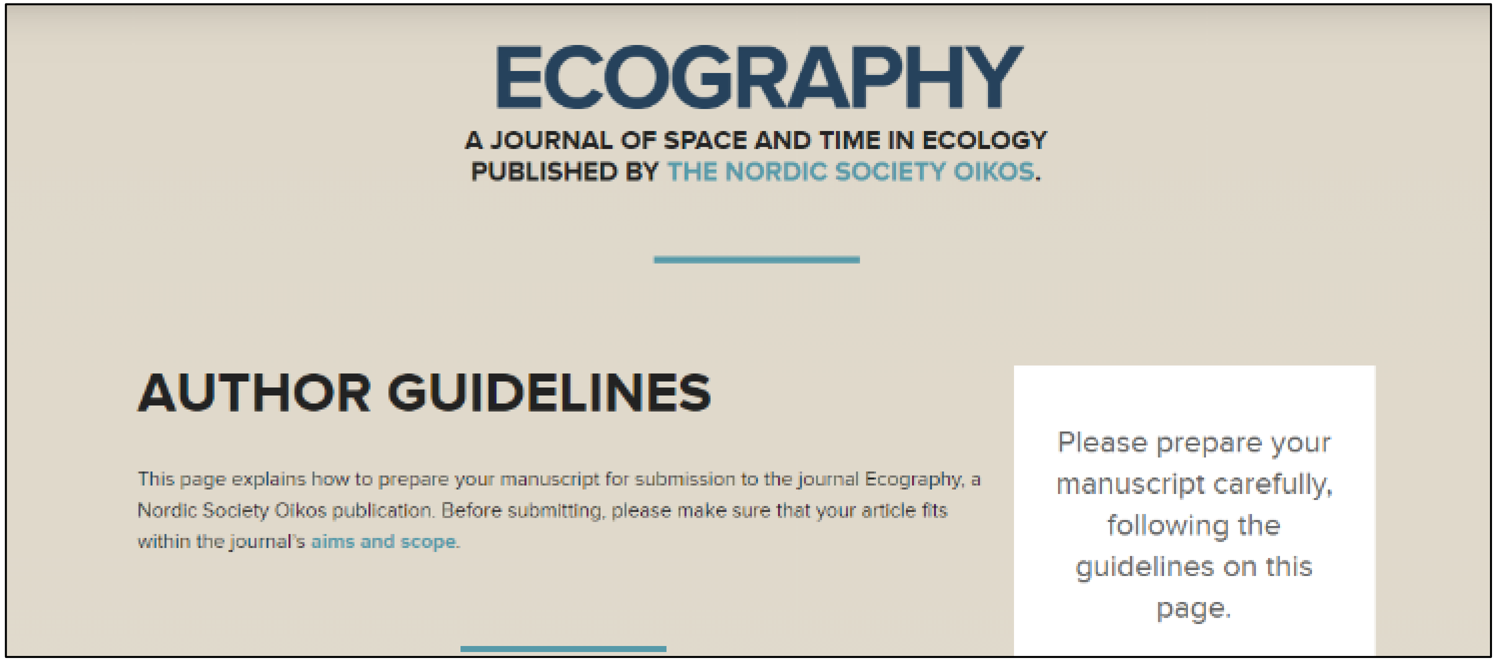
Quality control checks

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



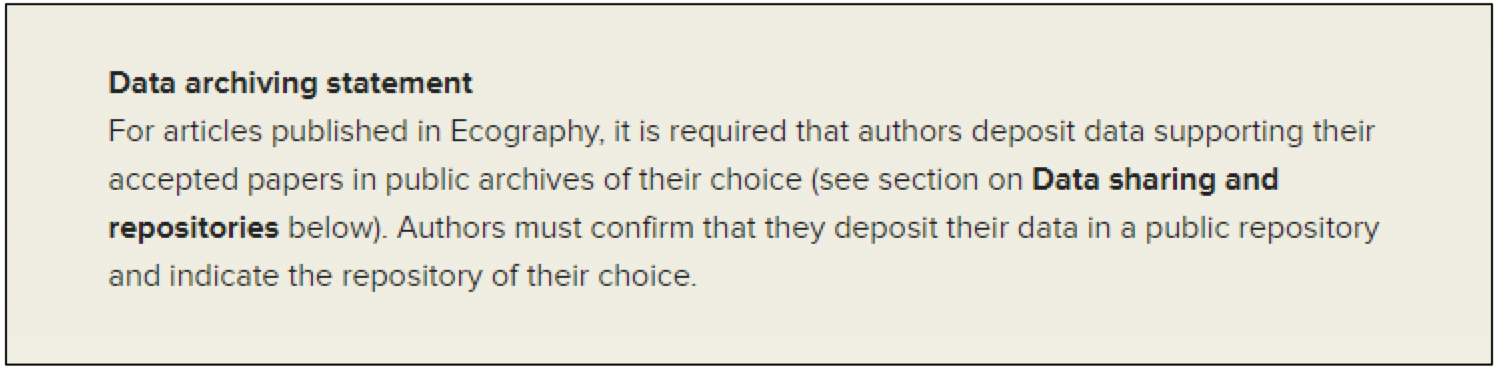
ECOGRAPHY

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.

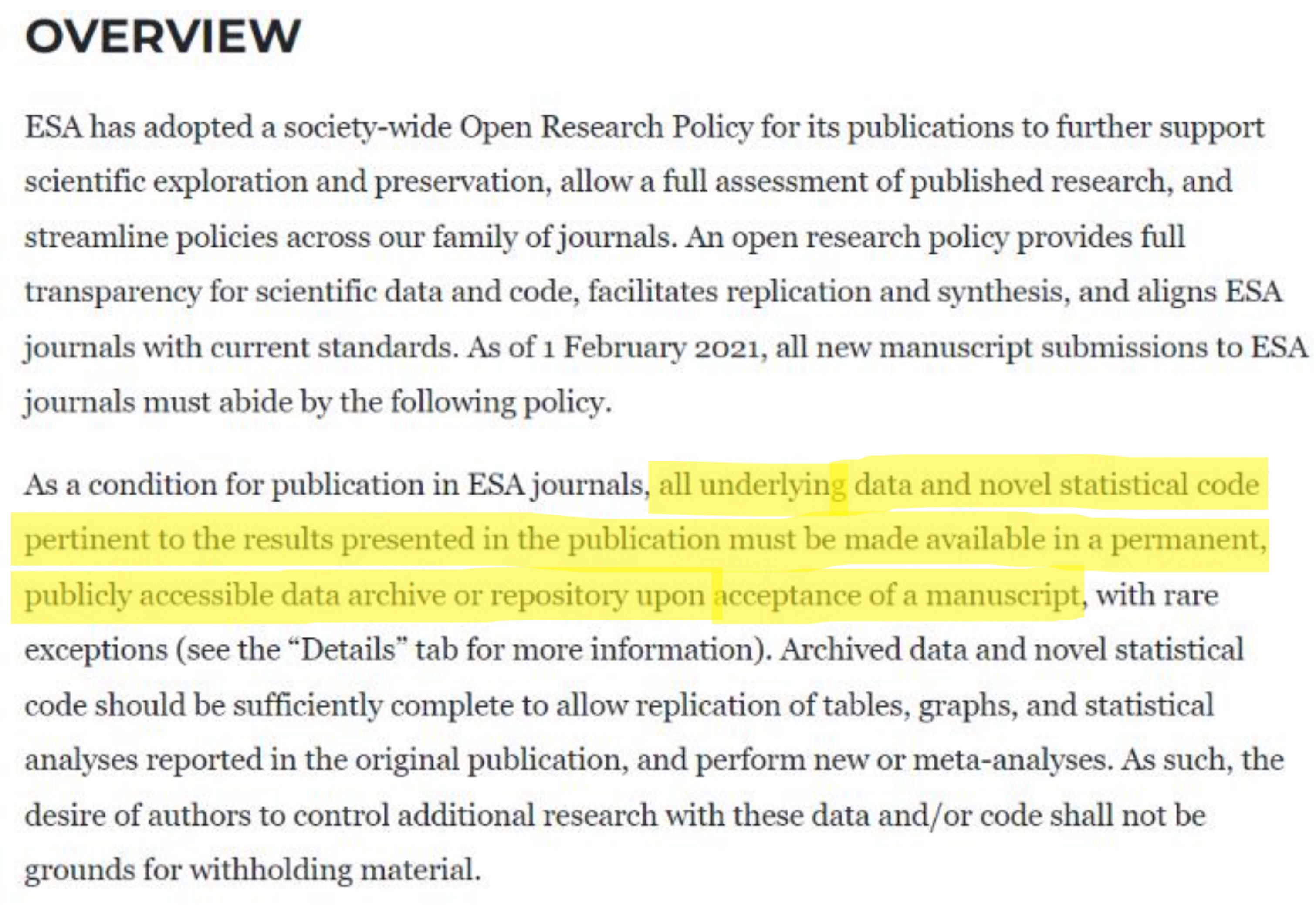
Please prepare your manuscript carefully, following the guidelines on this page.



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.

ECOLOGICAL SOCIETY OF AMERICA

esa

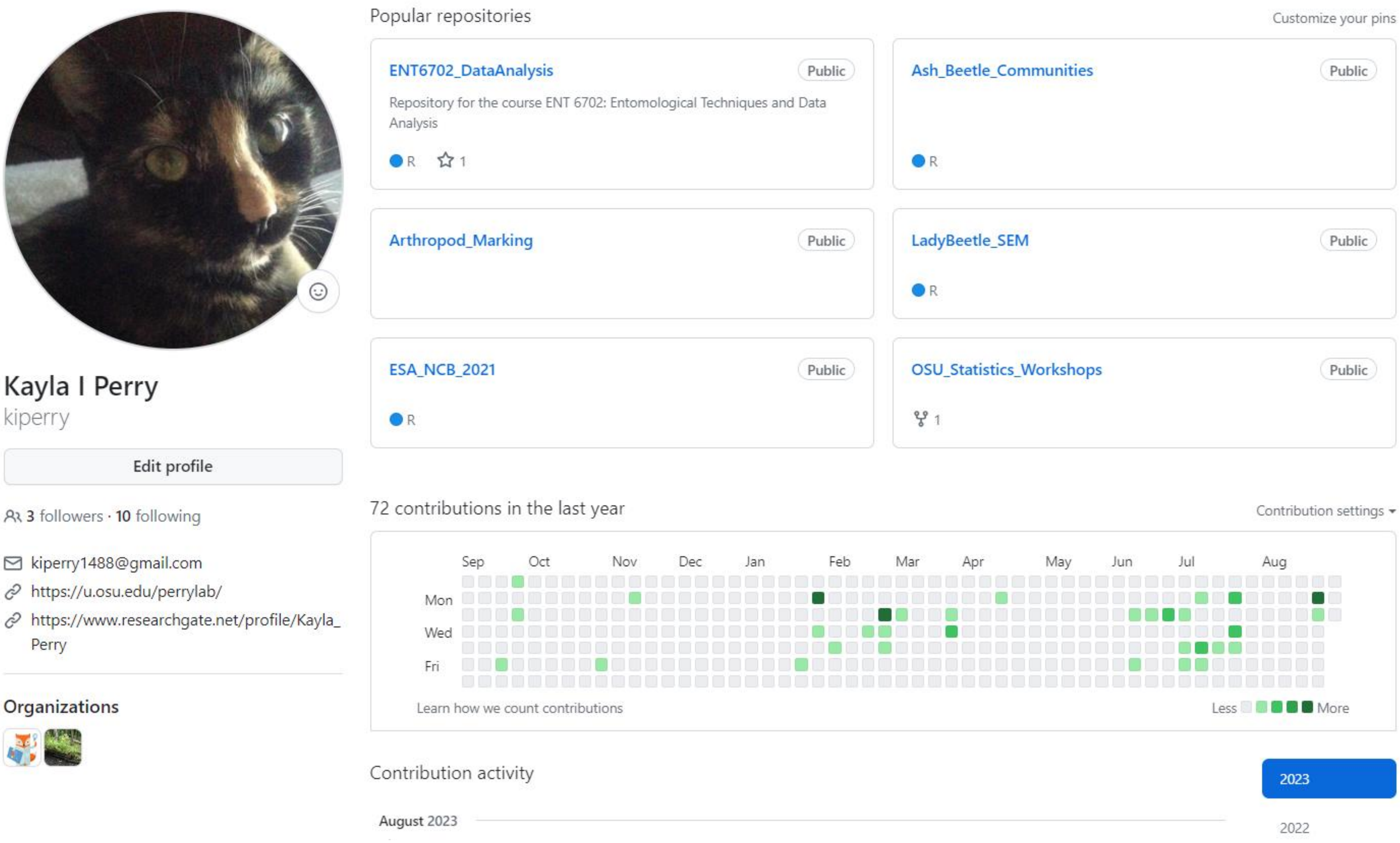
Git and GitHub – What are they?

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 ▾

Learn how we count contributions

Contribution activity

2023

2022

August 2023

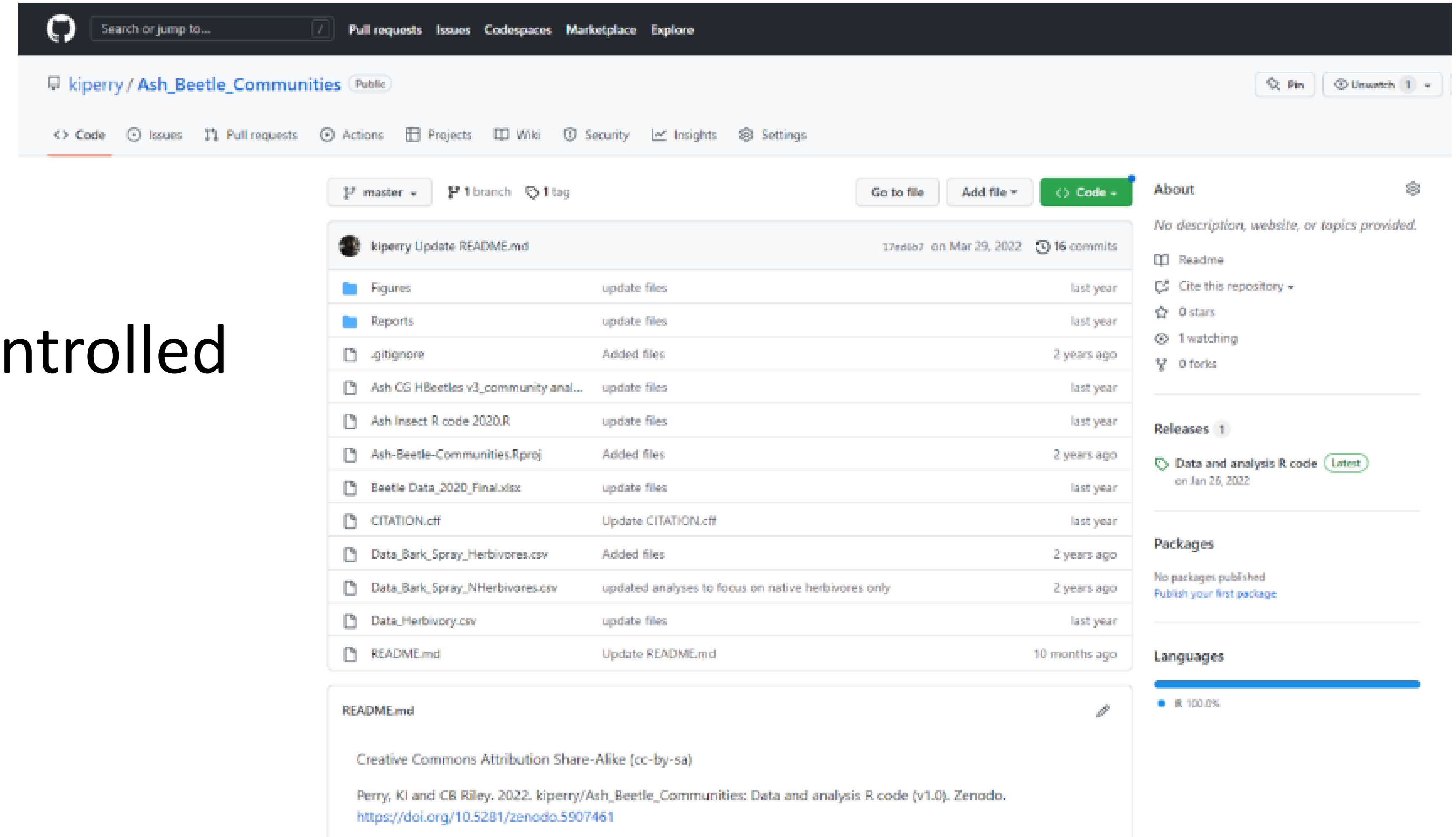
What is a repository (or repo)?

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



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

About

No description, website, or topics provided.

Readme

Cite this repository

0 stars

1 watching

0 forks

Releases 1

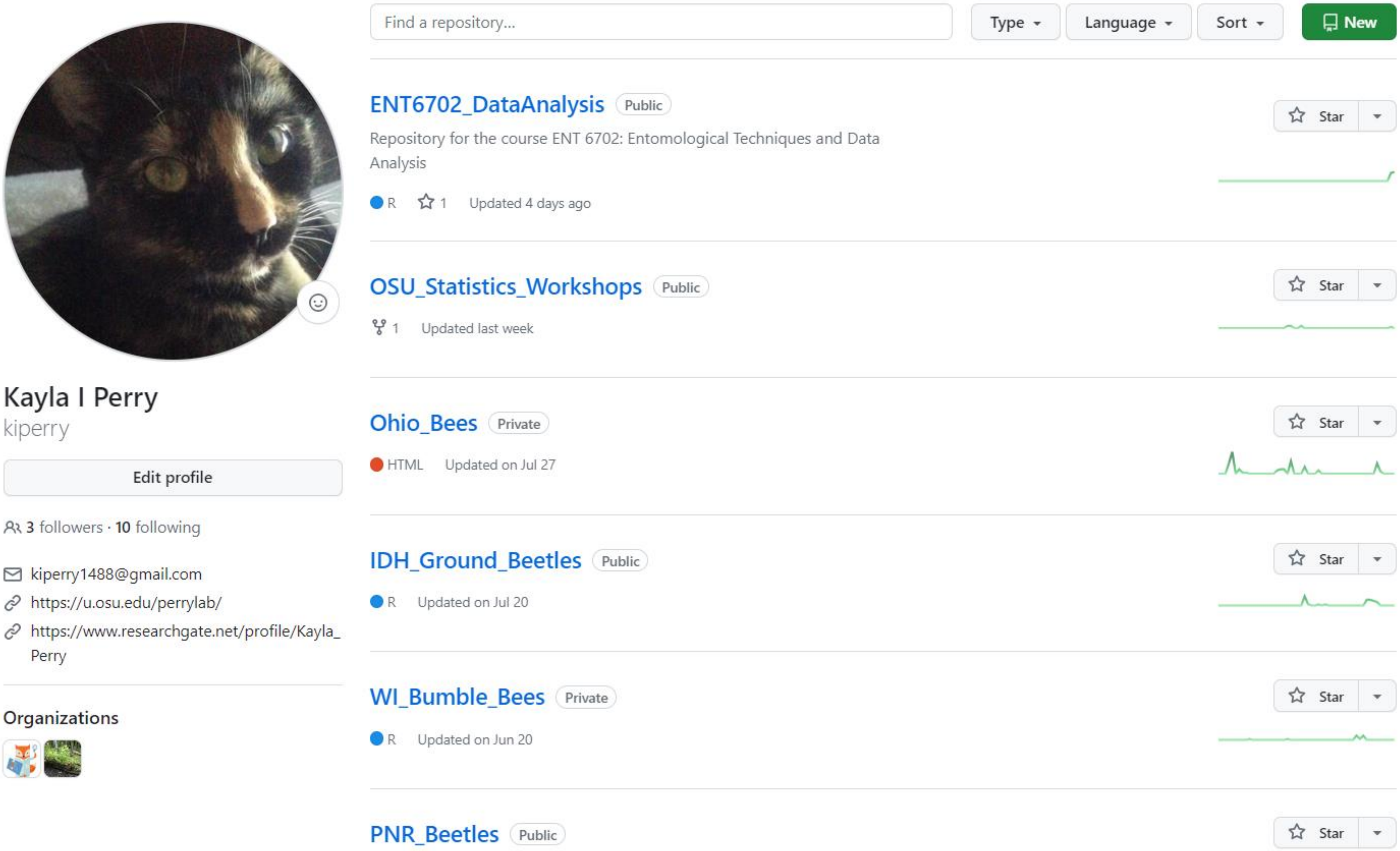
Data and analysis R code Latest
on Jan 26, 2022

Packages

No packages published
Publish your first package

Languages

● R 100.0%



Kayla I Perry
kiperry

3 followers · 10 following

kiperry1488@gmail.com
https://u.osu.edu/perrylab/
https://www.researchgate.net/profile/Kayla_Perry

ENT6702_DataAnalysis Public
Repository for the course ENT 6702: Entomological Techniques and Data Analysis
● R ☆ 1 Updated 4 days ago

OSU_Statistics_Workshops Public
1 Updated last week

Ohio_Bees Private
● HTML Updated on Jul 27

IDH_Ground_Beetles Public
● R Updated on Jul 20

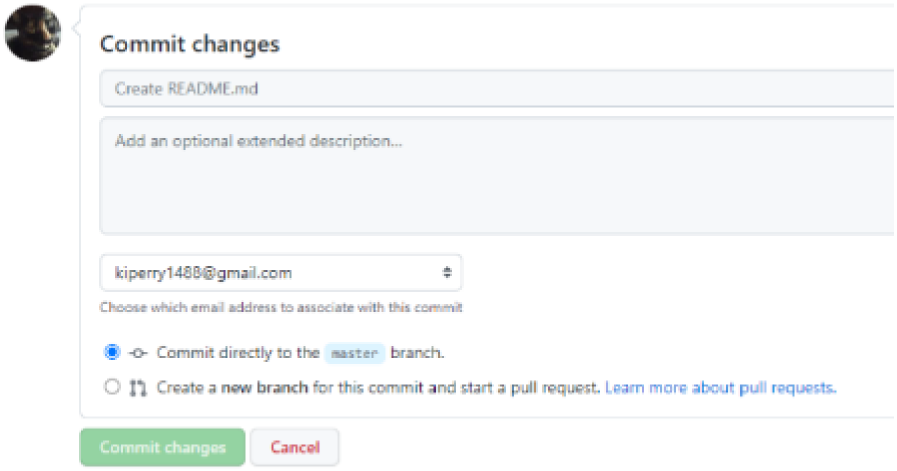
WI_Bumble_Bees Private
● R Updated on Jun 20

PNR_Beetles Public

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

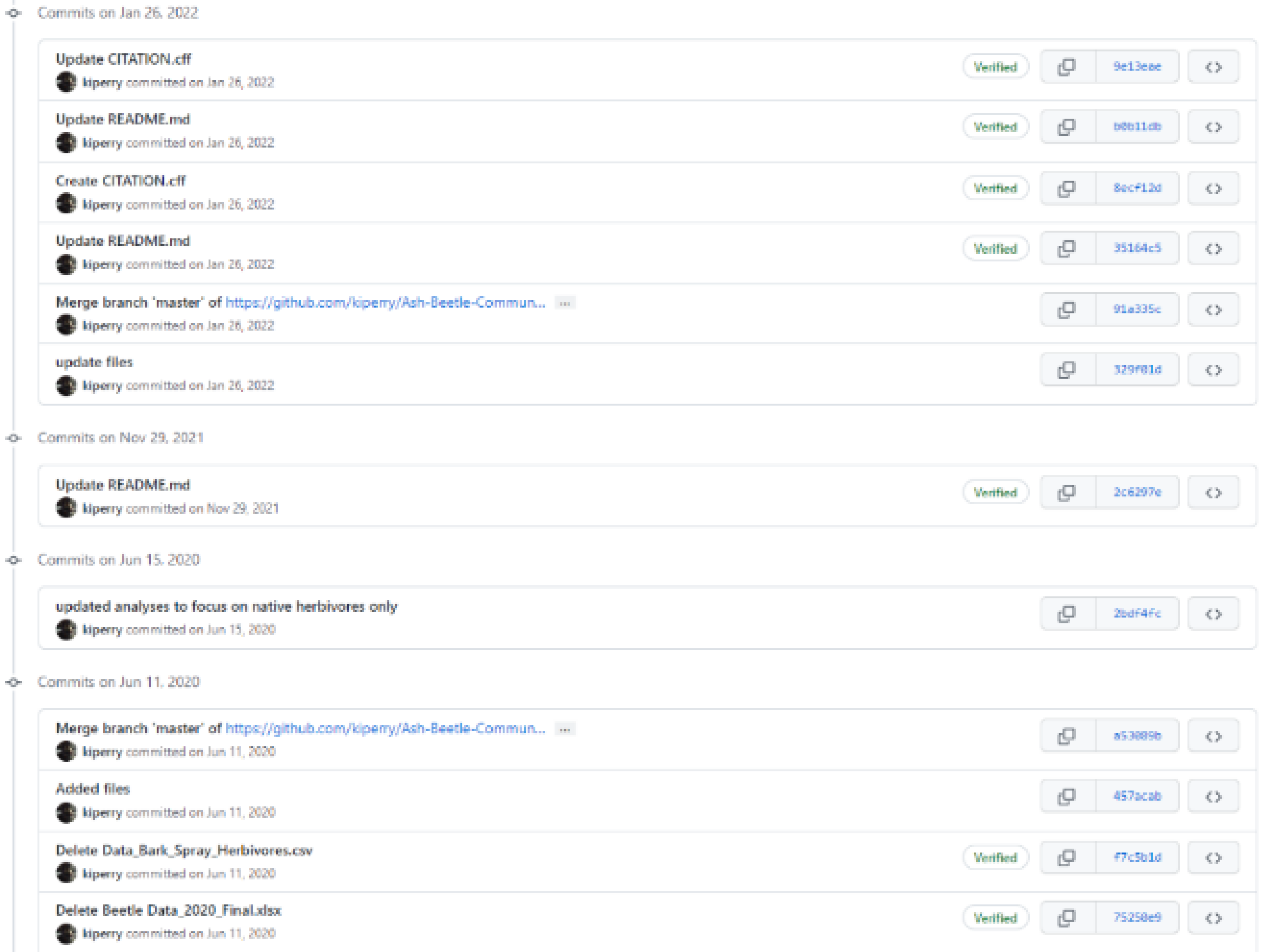
Add an optional extended description...

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.



Commits on Jan 26, 2022

- Update CITATION.cff — kiperry committed on Jan 26, 2022 — Verified — 9e13eae
- 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 — 329f01d

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 — 2bdf4fc

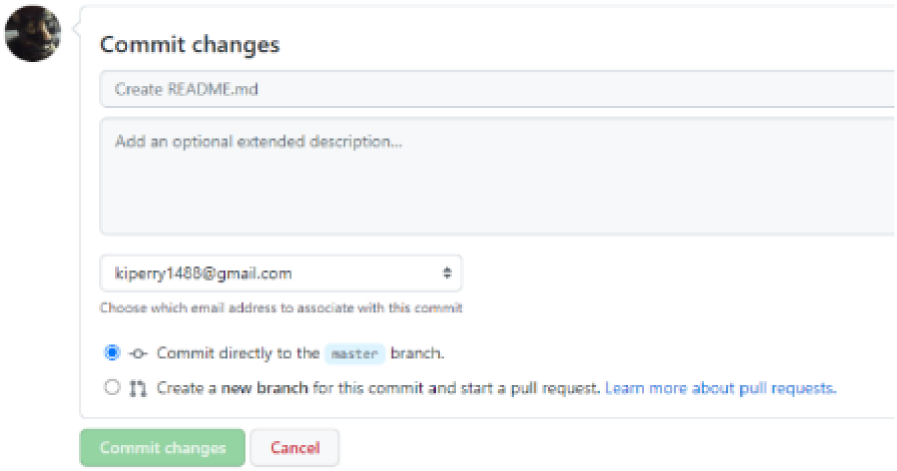
Commits on Jun 11, 2020

- Merge branch 'master' of https://github.com/kiperry/Ash-Beetle-Commun... … — kiperry committed on Jun 11, 2020 — a53089b
- Added files — kiperry committed on Jun 11, 2020 — 457acab
- Delete Data_Bark_Spray_Herbivores.csv — kiperry committed on Jun 11, 2020 — Verified — f7c5b1d
- Delete Beetle Data_2020_Final.xlsx — kiperry committed on Jun 11, 2020 — Verified — 75258e9

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

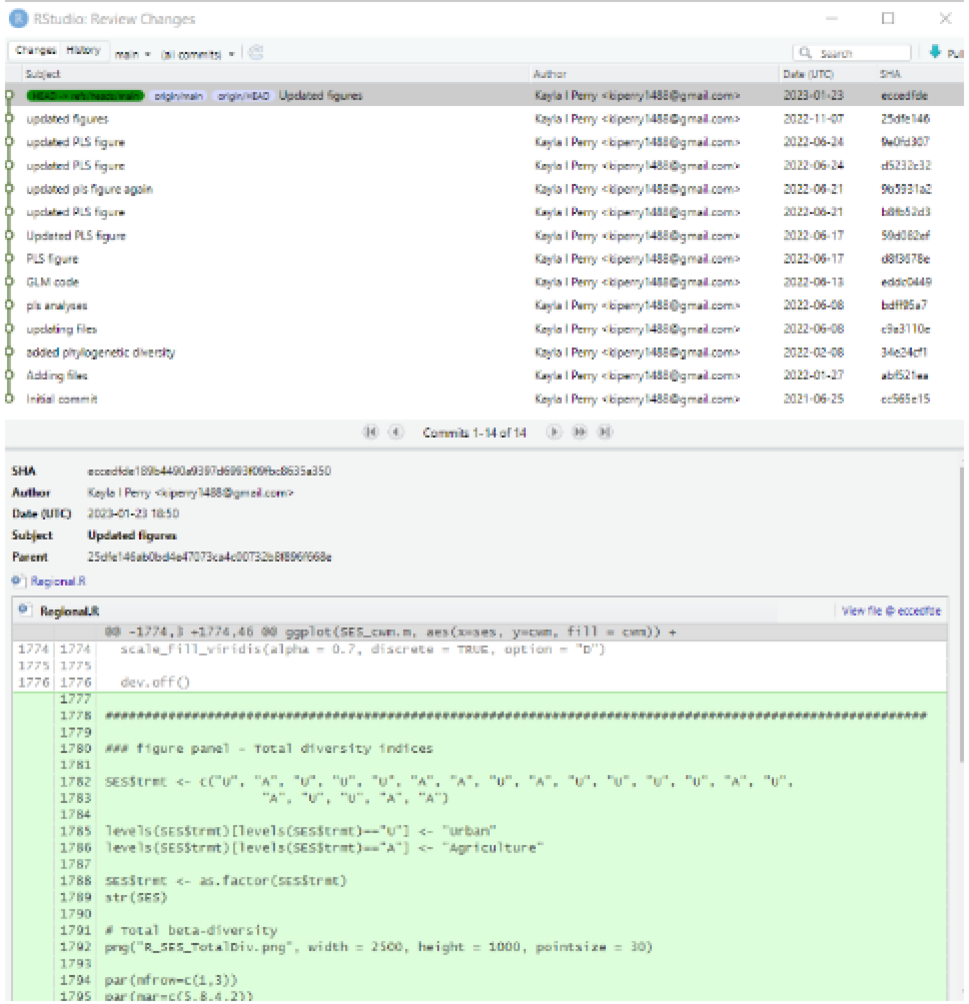
Add an optional extended description...

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.



Subject	Author	Date (UTC)	SHA
HEAD -> refs/heads/main origin/main origin/HEAD Updated figures	Kayla I Perry <kiperry1488@gmail.com>	2023-01-23	eccedfde
updated figures	Kayla I Perry <kiperry1488@gmail.com>	2022-11-07	25dfe146
updated PLS figure	Kayla I Perry <kiperry1488@gmail.com>	2022-06-24	9e0fd307
updated PLS figure	Kayla I Perry <kiperry1488@gmail.com>	2022-06-24	d5232c32
updated pls figure again	Kayla I Perry <kiperry1488@gmail.com>	2022-06-21	9b5931a2
updated PLS figure	Kayla I Perry <kiperry1488@gmail.com>	2022-06-21	b8fb52d3
Updated PLS figure	Kayla I Perry <kiperry1488@gmail.com>	2022-06-17	59d082ef
PLS figure	Kayla I Perry <kiperry1488@gmail.com>	2022-06-17	d8f3678e
GLM code	Kayla I Perry <kiperry1488@gmail.com>	2022-06-13	eddc0449
pls analyses	Kayla I Perry <kiperry1488@gmail.com>	2022-06-08	bdff95a7
updating Files	Kayla I Perry <kiperry1488@gmail.com>	2022-06-08	c9a3110e
added phylogenetic diversity	Kayla I Perry <kiperry1488@gmail.com>	2022-02-08	34e24cf1
Adding files	Kayla I Perry <kiperry1488@gmail.com>	2022-01-27	abf521ea
Initial commit	Kayla I Perry <kiperry1488@gmail.com>	2021-06-25	cc565e15

Commits 1-14 of 14

SHA eccedfde189b4490a9397d6993f09fbc8635a350
Author Kayla I Perry <kiperry1488@gmail.com>
Date (UTC) 2023-01-23 18:50
Subject **Updated figures**
Parent 25dfe146ab0bd4e47073ca4c00732b8f8969068e

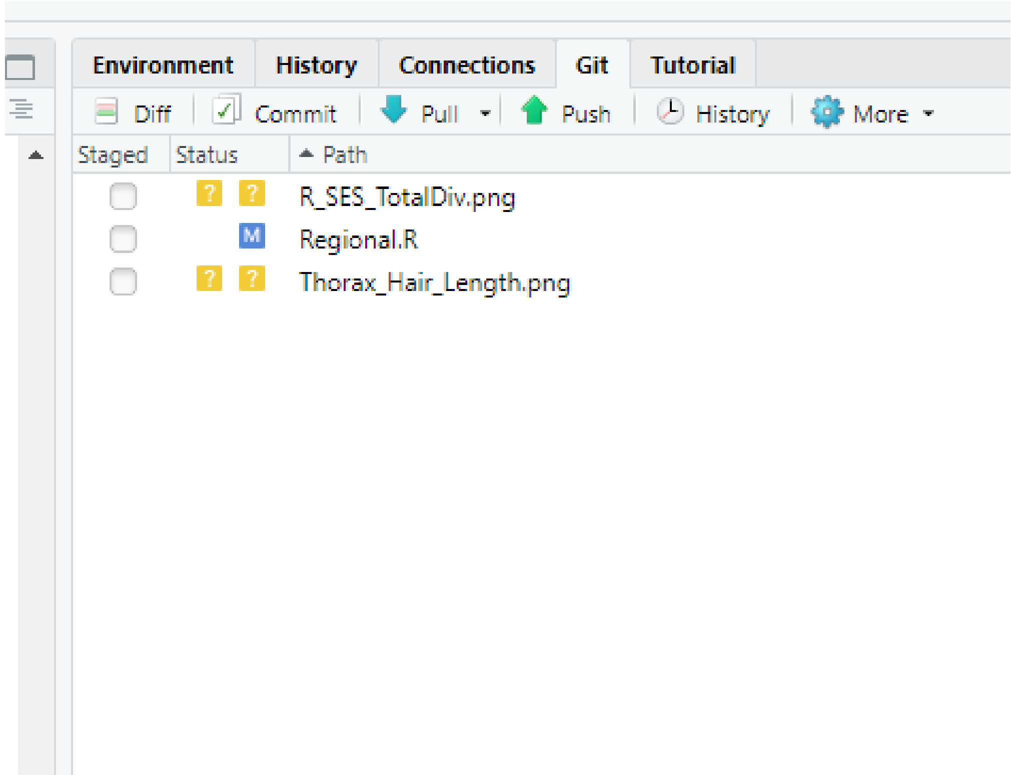
Regional.R

Regional.R View file @ eccedfde

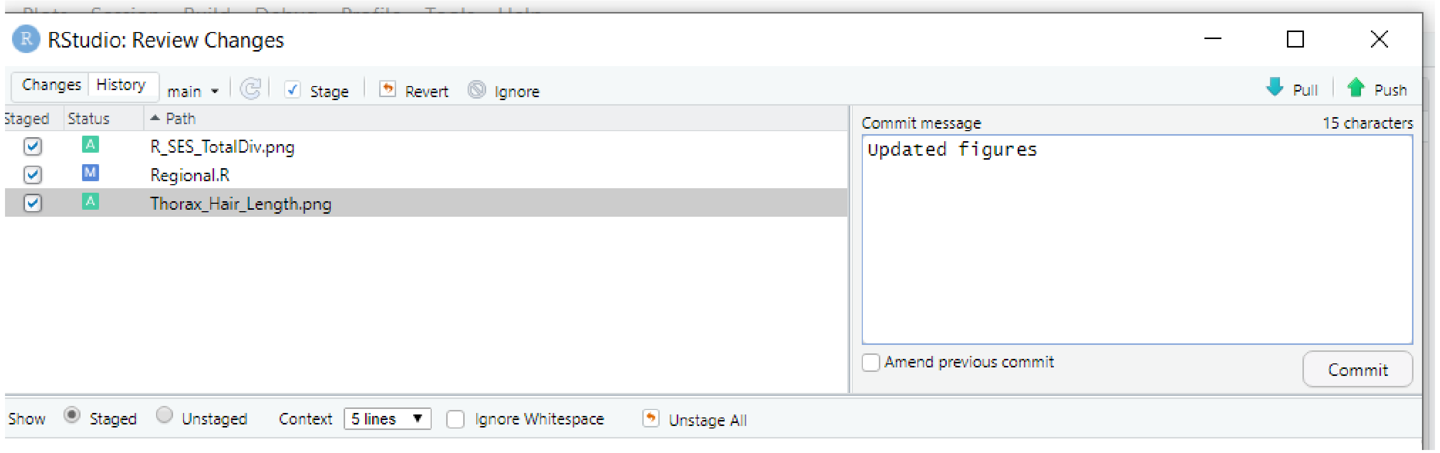
```
@@ -1774,3 +1774,46 @@ ggplot(SES_cwm.m, aes(x=ses, y=cwm, fill = cwm)) +
  scale_fill_viridis(alpha = 0.7, discrete = TRUE, option = "D")

  dev.off()

#####################################################################################################

### figure panel - Total diversity indices

SES$trmt <- c("U", "A", "U", "U", "U", "A", "A", "U", "A", "U", "U", "U", "U", "A", "U",
                  "A", "U", "U", "A", "A")

levels(SES$trmt)[levels(SES$trmt)=="U"] <- "Urban"
levels(SES$trmt)[levels(SES$trmt)=="A"] <- "Agriculture"

SES$trmt <- as.factor(SES$trmt)
str(SES)

# Total beta-diversity
png("R_SES_TotalDiv.png", width = 2500, height = 1000, pointsize = 30)

par(mfrow=c(1,3))
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



Staged	Status	Path
□	? ?	R_SES_TotalDiv.png
□	M	Regional.R
□	? ?	Thorax_Hair_Length.png



Staged	Status	Path
☑	A	R_SES_TotalDiv.png
☑	M	Regional.R
☑	A	Thorax_Hair_Length.png

Commit message — 15 characters

```
Updated figures
```

□ Amend previous commit

Show ● Staged ○ Unstaged Context 5 lines □ Ignore Whitespace Unstage All

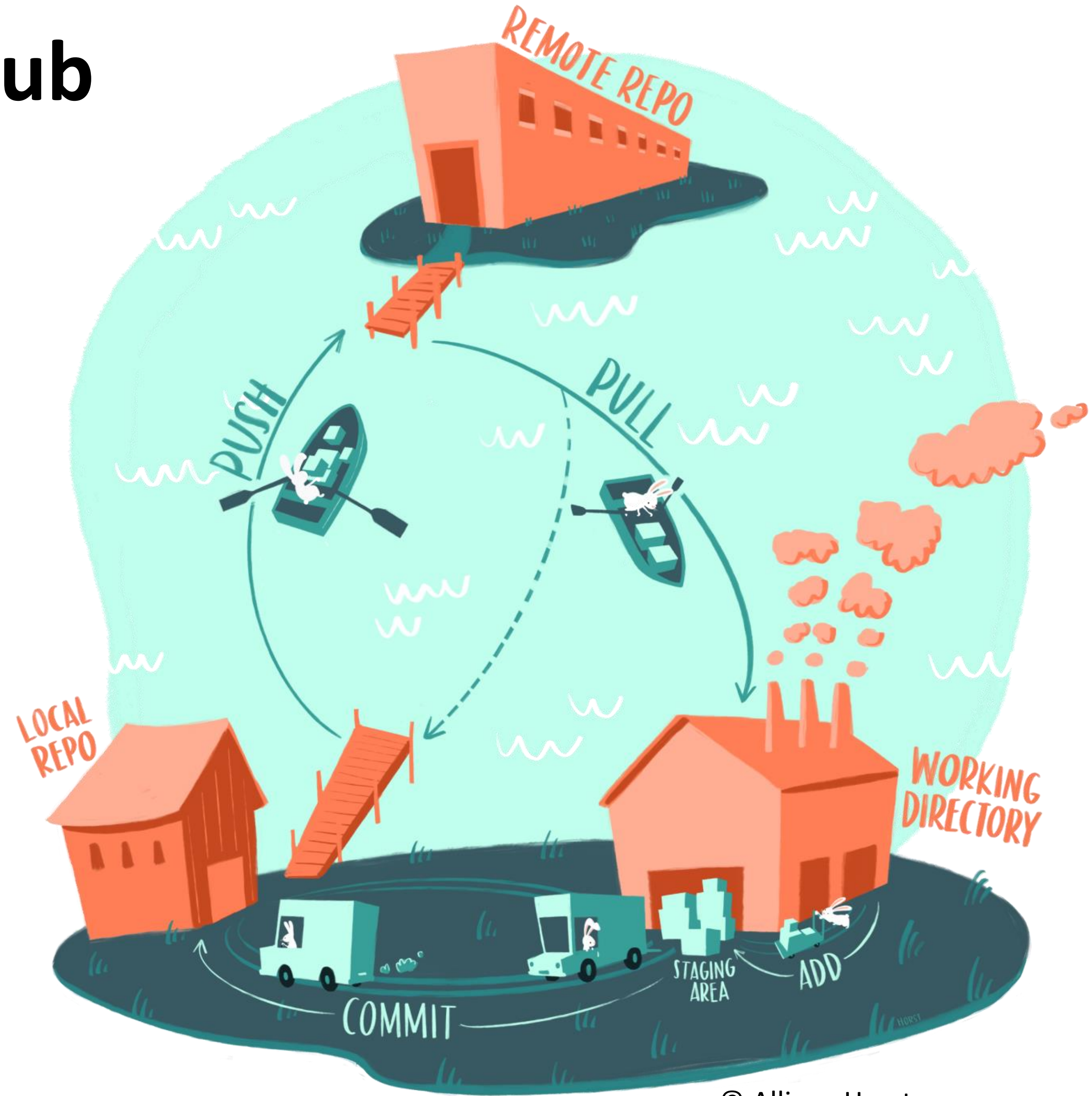
1) Sync project files locally on your computer and online

2) Make commits to record changes to files over time

3) Facilitates remote collaboration within a project repository

4) Promotes open science

Git and GitHub



© Allison Horst

What can we do with Git/GitHub?

1) Experiment on projects without breaking them – **Branch**

2) Make, assign, and keep track of tasks – **Issues**

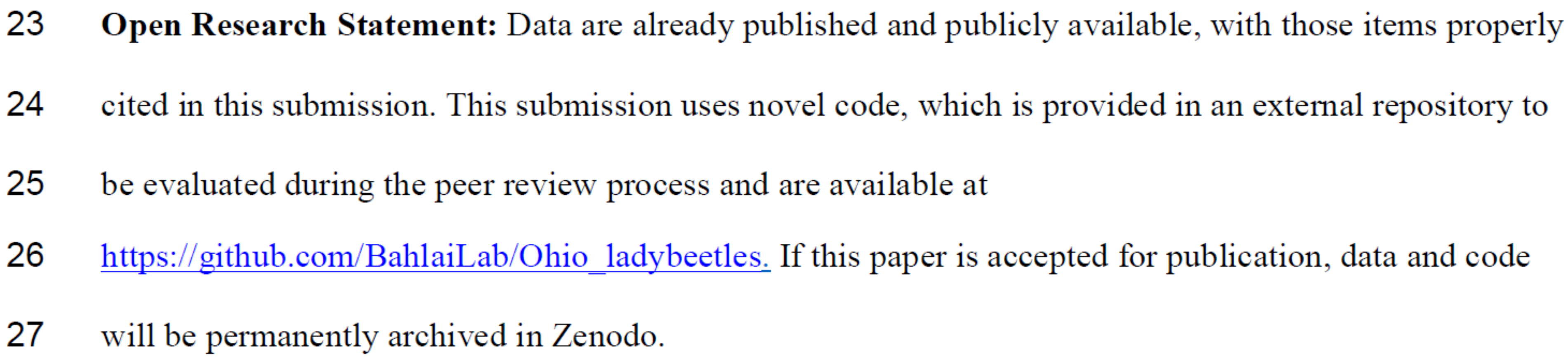
3) Access existing projects made by others – **Fork or Clone**

4) Build on existing projects with collaborators – **Pull, Commit, Push**

Submitting your manuscript for review?

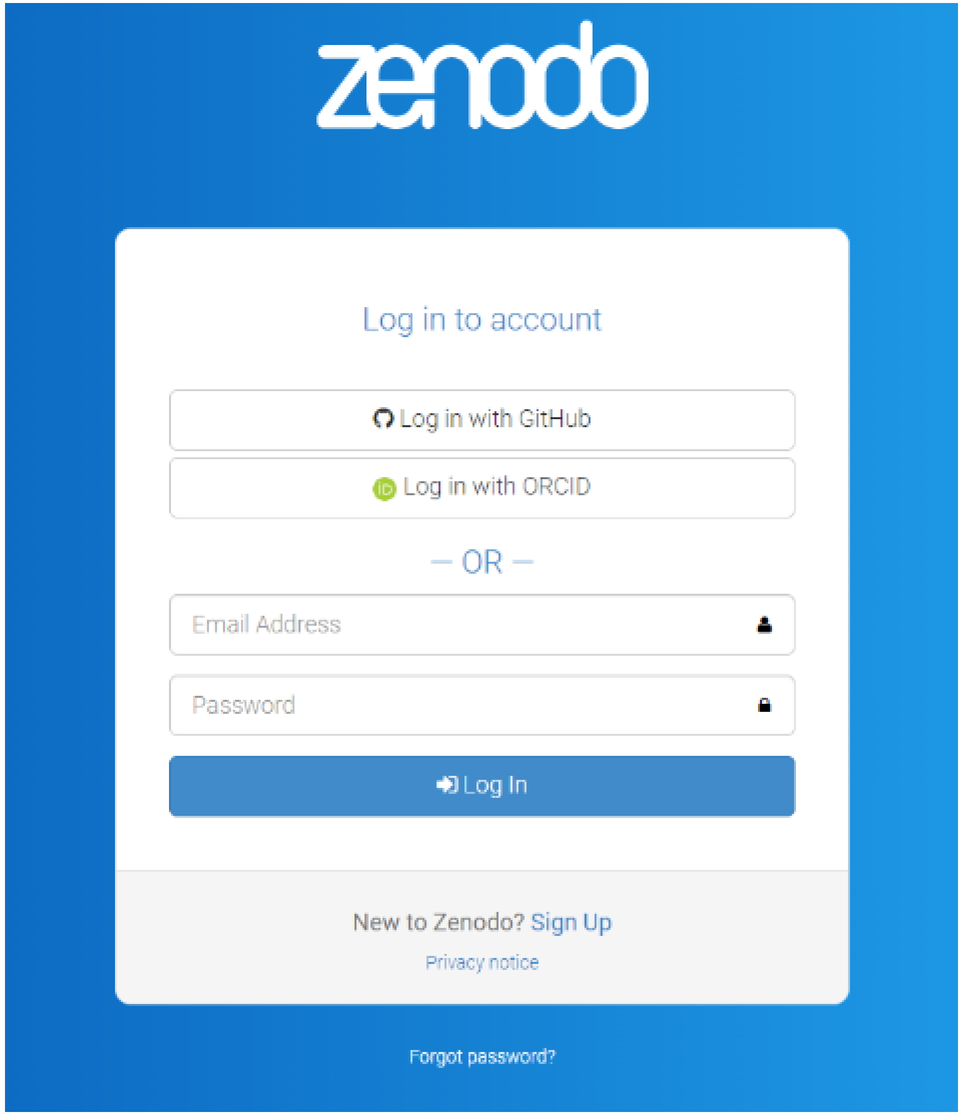
Include your GitHub repository link in your open research/data availability statement

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



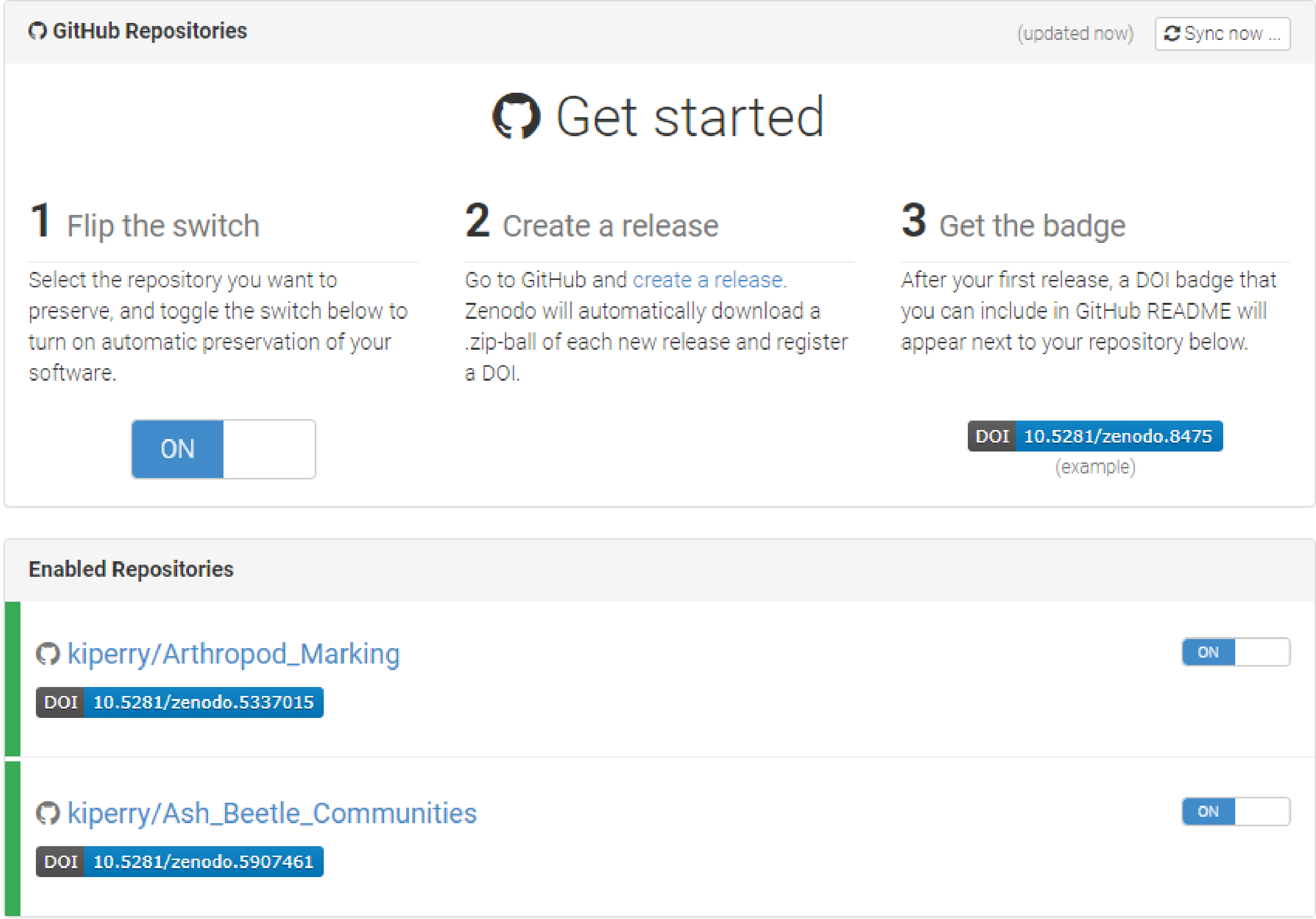
Open Research Statement: Data are already published and publicly available, with those items properly cited in this submission. This submission uses novel code, which is provided in an external repository to be evaluated during the peer review process and are available at https://github.com/BahlaiLab/Ohio_ladybeetles.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

Link GitHub repository with Zenodo for DOI



Developed under European OpenAIRE Program

Operated by CERN



1 Flip the switch

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

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.

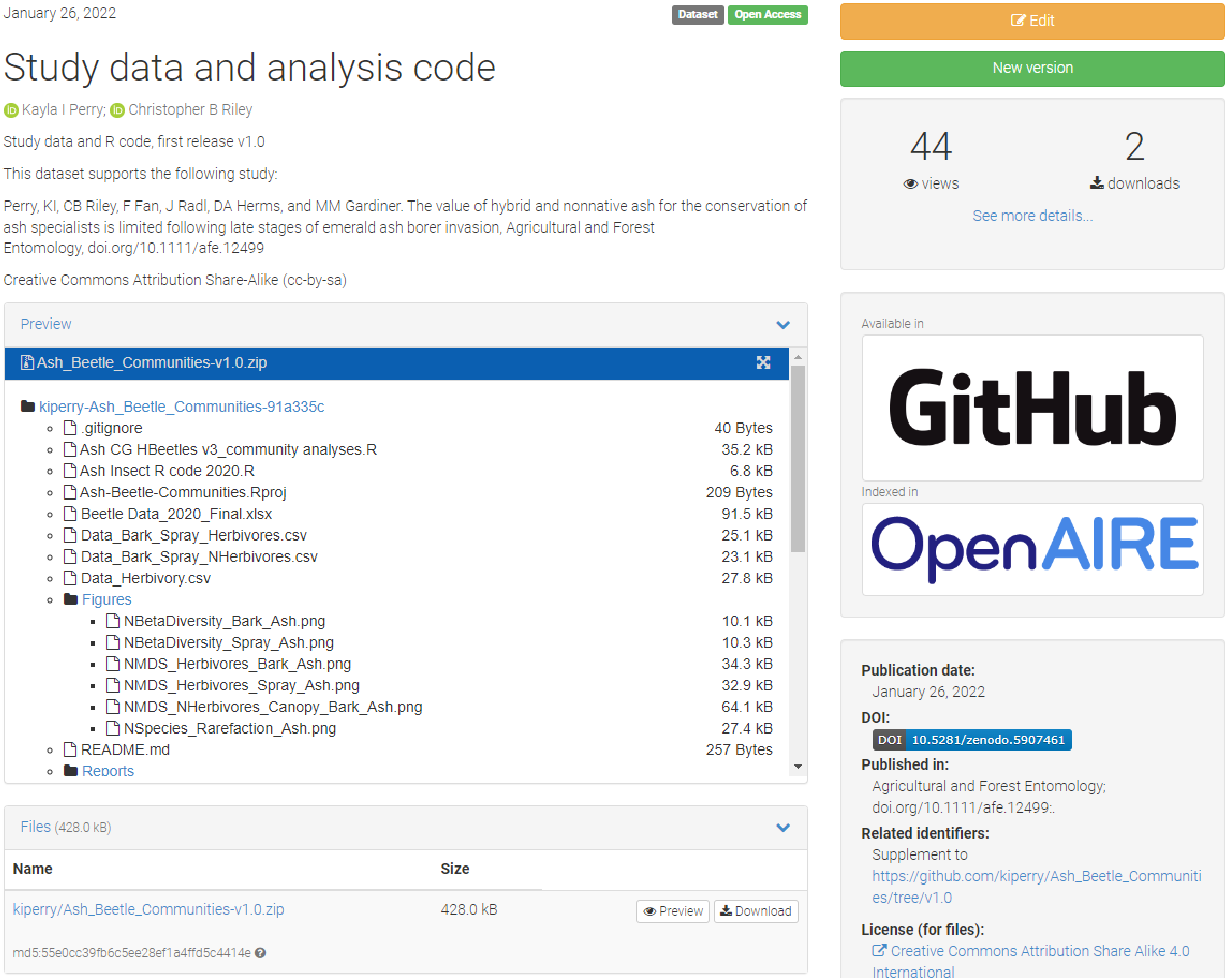
DOI 10.5281/zenodo.8475 (example)

Enabled Repositories

kiperry/Arthropod_Marking — DOI 10.5281/zenodo.5337015 — ON

kiperry/Ash_Beetle_Communities — DOI 10.5281/zenodo.5907461 — ON

Software and Data Products category on your CV!



January 26, 2022

Dataset Open Access

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 md5:55e0cc39fb6c5ee28ef1a4ffd5c4414e	428.0 kB	Preview Download

44 views

2 downloads

See more details...

Available in

GitHub

Indexed in

OpenAIRE

Publication date:
January 26, 2022

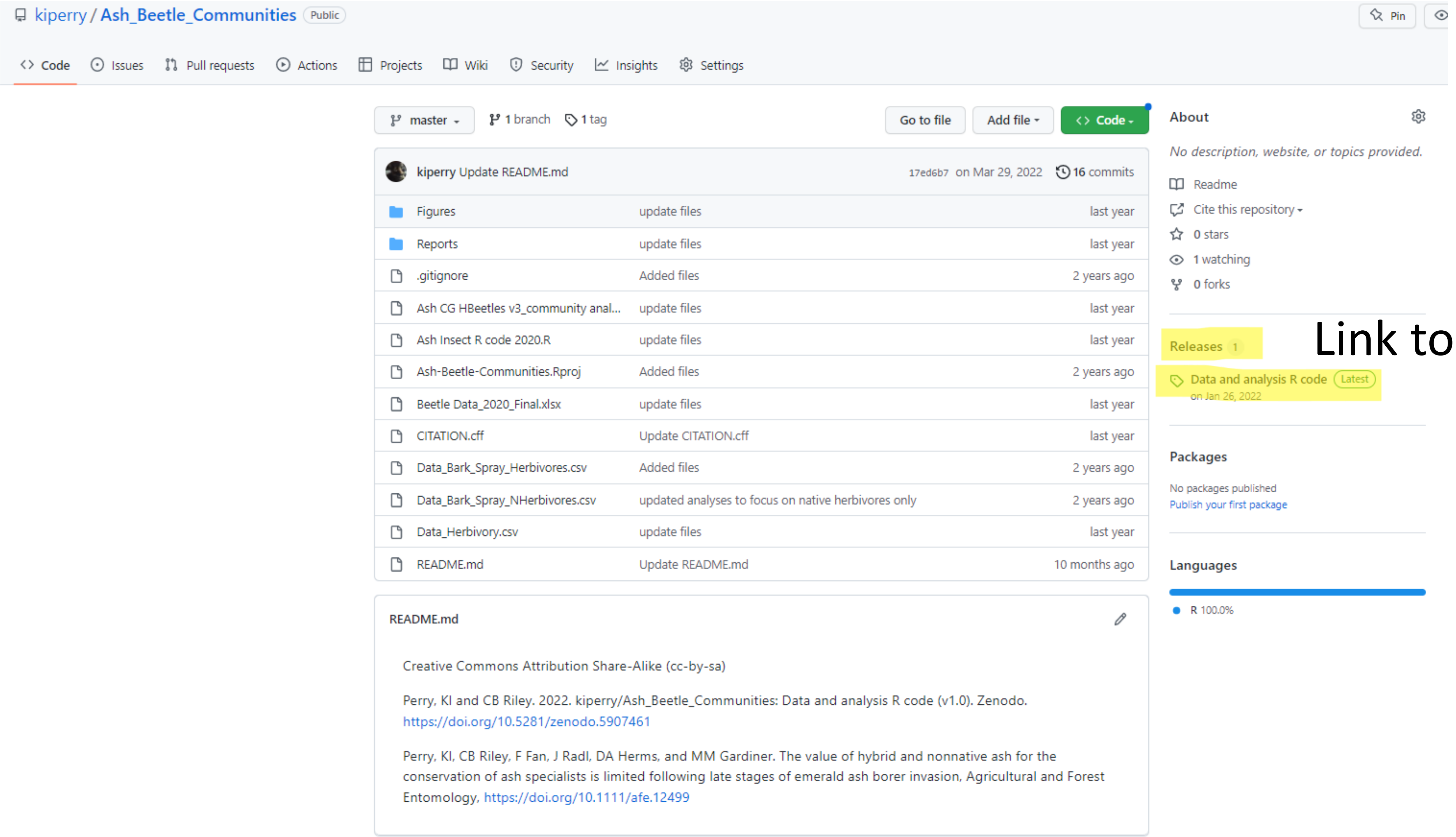
DOI:
DOI 10.5281/zenodo.5907461

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 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

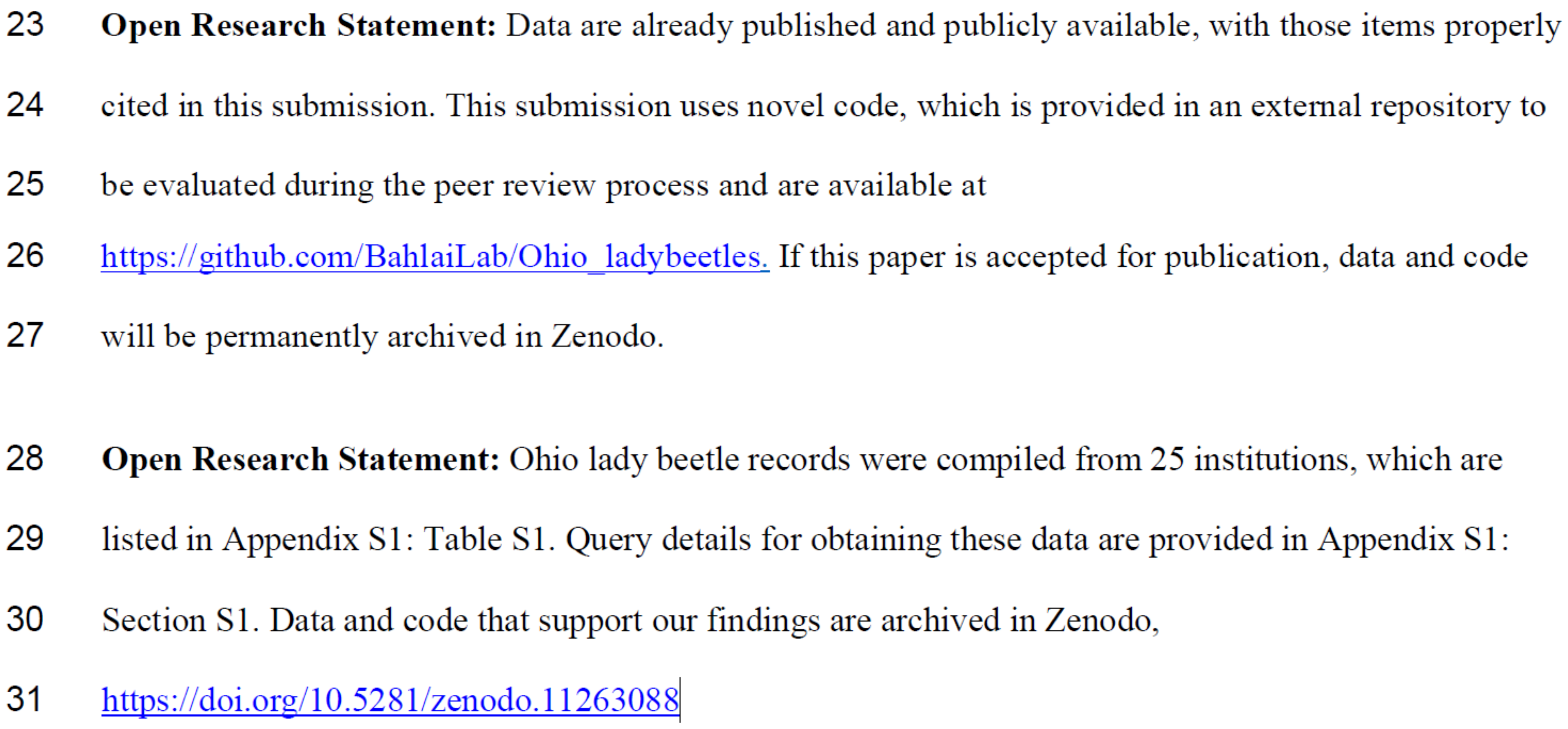
Languages

R 100.0%

Link to Zenodo

Submitting your manuscript for review?

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



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

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