Not just for programmers: A friendly guide on the versatility/benefits of GitHub for accelerating collaborative research in Ecology and Evolution

This manuscript (permalink) was automatically generated from SORTEE-Github-Hackathon/manuscript@ab6f9c9 on January 25, 2022.

Authors

- **Hermione Jean Granger**
  XXXX-XXXX-XXXX-XXXX · hermionejeangranger · hermionejeangranger
  Department for the Regulation and Control of Magical Creatures, British Ministry of Magic, United Kingdom; Department of Magical Law Enforcement, British Ministry of Magic, United Kingdom · Funded by Hogwarts Grant XXXXXX-XX.

- **Dylan G. E. Gomes**
  0000-0002-2642-3728 · dylangomes
  Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, Newport, OR, United States

- **Cole B. Brookson**
  0000-0003-1237-4096 · colebrookson
  Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

- **Robert Crystal-Ornelas**
  0000-0002-6339-1139 · robcrystalornelas · rob_c_ornelas
  Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

- **Ali Guncan**
  0000-0003-1765-648X · Aguncan · aliguncan
  Department of Plant Protection, Faculty of Agriculture, Ordu University, 52200, Ordu, Turkey

- **Brandon P.M. Edwards**
  0000-0003-0865-3076 · BrandonEdwards
  Department of Biology, Carleton University, Ottawa, ON K1S 5B6, Canada

- **Kaitlyn M. Gaynor**
  0000-0002-5747-0543 · kaitlyngaynor · kaitlyngaynor
  Departments of Zoology and Botany, University of British Columbia, Vancouver, BC, Canada; National Center for Ecological Analysis and Synthesis, Santa Barbara, CA 93101, USA

- **Katherine Hébert**
  0000-0001-7866-6775 · katherinehebert · hebert_kat
  Département de biologie, Université de Sherbrooke, Québec, Canada

Abstract

Importance

Introduction

High-level/general background about GitHub

contributors to this section: RCO,

With over 73 million registered users, GitHub and its underlying version control system Git, are the defacto platform for collaboration on computer code [1]. GitHub has become an indispensible tool for software developers because,through version control, users can track changes to multiple files and folders over time [2]. Thus, users have an "audit trail" on the files they choose to store on their GitHub repositroies which simultanously less "ad hoc" than passing files back and forth yet able to scale up as projects take on more files or collaborators [3].

Researchers in ecology and evolutionary biology (EEB) are starting to collaborate on software as part of their research, and some are interacting with GitHub for the first time [4]. For first-time users, the GitHub learning curve can seem overwhelming because the platform and its features were created with collaboration on software as the central goal.
However, by leveraging existing tools on GitHub and the wide range of collaborations they can enable, researchers in EEB can make the most out of their research and collaborative projects.

What's already been written about GitHub

contributors to this section: RCO, Git is the version control system that enables all the collaborative tools available on GitHub. Because the details of interacting with Git can get very technical very quick, we focus instead on the web platform GitHub. However, we suggest those interested in Git explore the many papers [5] [6] and books [7/] that can provide an introduction to git. Despite the prevelance of technical papers and books that focus on Git or GitHub for the software development community, there are much fewer resources for EEB researchers who want to begin collaborating through GitHub. We acknowledge that GitHub is not the only way for productive collaboration on cloud-based research documents and code, so we encourage researchers in EEB to take the elements of a GitHub that fit into their workflow. Because GitHub as web platform is so well documented and has a robust user community, scientists can take advantage of many collaborative aspects without knowing even a line of Git code.

What's already been done with GitHub in EcoEvo

contributors to this section:

Very friendly description of what GitHub is and the main uses and advantages of using it in the natural sciences back in 2016 (Perkel 2016) What's missing about GitHub in EcoEvo and our objective: Introducing the GitHub ecosystem that's composed of many different elements!

What's missing about GitHub in EcoEvo and our objective

contributors to this section:

Simple habits (of which github is one component) can do a lot to make research more reproducible and collaborative (Alston and Rick 2021). In EcoEvo Github use is predicated on an understanding in R. This close connection has some benefits, but other programming languages are frequently used by researchers (e.g. Python, Julia). Lots of ways to use GitHub that are independent from R. We have in

this hackathon a definite focus on R tools for interacting with GitHub, but sometimes the issues we present at 'Github' issues might be more about the ways that we interact with Github (i.e. through R vs. bash shell)

GitHub in EcoEvo examples (Part 1)

Storing and archiving version-controlled data

contributors to this section: Another potential use case/user perspective: Some people are just using GitHub to backup their data, use their code on differnt machines. Just push and pull (Box 1) from their own repo. GitHub integrates with Zenodo, a popular, free data archiving service funded by CERN. After linking your GitHub account to Zenodo and turning on archiving, any time a release (Box 1) is made, a snapshot of the entire repository is archived in Zenodo with a versioned, citable DOI

Virtual lab notebook

contributors to this section: commits as a way to record daily progress issues as a way to keep track of short-term objectives/goals, and progress towards them

Responding to reviewer comments

contributors to this section: using github issues (Box 1) to organize and respond to reviewer comments on a manuscript. See example here

Classroom teaching / educational materials

contributors to this section: Matthew D. Beckman, Mine Çetinkaya-Rundel, Nicholas J. Horton, Colin W. Rundel, Adam J. Sullivan & Maria Tackett (2021) Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses, Journal of Statistics and Data Science Education, 29:sup1, S132-S144, DOI: 10.1080/10691898.2020.1848485

GitHub in EcoEvo examples (Part 2)

Project management

Contributors to this section: Kaitlyn Gaynor, Rob Crystal-Ornelas

GitHub can be a powerful tool for team-based project management, allowing collaborators to share feedback, brainstorm ideas, and troubleshoot problems. The "Issues" feature of GitHub allows for discrete tasks and sub-tasks to be identified, assigned to team members, and categorized with custom labels, and the new "Discussion" feature serves as a message board for conversation. Scripts, commit messages, and pull requests can be linked directly to issues and discussions, providing a clear record of project workflow. The use of GitHub for all project-related conversation and planning, rather than e-mail or messaging tools, makes it easier to keep track of progress throughout the lifespan of a project and less likely for issues to slip through the cracks. It is not essential for all team members to have proficiency in git or programming, as users can interact with Issues and Discussions via web browser or e-mail. By default, GitHub repositories are publicly visible, and so anyone with a GitHub account can not only view content, but also engage with repository administrators through Issues and Discussions.

Can talk about ESS-DIVE's project management using ZenHub/Jira to manage customer support requests, feature updates to our data sharing platform.

Building website

Contributors to this section:

Seems like the technical aspect of this is discussed in Dawson, Chris (2016). Building Tools with GitHub: Customize Your Workflow. O'Reilly Media GitHub pages allows any .html document to be rendered as a website with a URL. This could be, for example, a report written in markdown or R Markdown rendered into a .html file.

Making code citable

Contributors to this section: Rob Crystal-Ornelas
GitHub makes it easy to store and share a variety of data files in the cloud. But for a variety of reasons (e.g., privately owned company, ability to make repositories private, accounts can be deleted at will) GitHub is not considered a long-term data or code repository like zenodo and figshare [8] [6.]. Also, unlike the long-term repositories, GitHub does not issue Digital Object Identifiers (DOIs) for content uploaded to their servers. DOIs are persistent and unique alpha-numeric IDs assigned to research products like papers, code, and data. DOIs allows tracking and citing research products. For this reason, scientists who share code and data through GitHub are strongly encouraged to also submit GitHub repository content to a long-term data archive [9]. Fortunately, both long-term repositories mentioned above (zenodo and figshare) have integrations with GitHub which facilitates archiving a snapshot of all repository content with the click of a button.

GitHub Linking with Zenodo, etc. to achieve a DOI helps work become findable, gives proper attribution (Hampton et al. 2015) Another key step researchers should consider taking when they receive a DOI for the content they archive on GitHub is choose a usage license [[10]. This helps

Collaborative (code) editing

Contributors to this section: Kaitlyn Gaynor, Rob Crystal-Ornelas

From its inception, one of the primary uses of GitHub has been for collaborative coding. We acknowledge that the average software developer and EcoEvo researcher using GitHub

Is it worth walking through how collaborative code editing works through GitHub, or just pointing to all the available resources for this? (e.g. https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/incorporating-changes-from-a-pull-request/merging-a-pull-request). Rob: I think pointing to available resources and citing them here is great [10].

GitHub can facilitate interactions between research advisors and advisees, providing a platform for students or other trainees to share in-progress code, and flag specific challenges or questions for their supervisors or mentors. Periodic code review can also help advisors to identify errors early in the process, and inform further training and mentorship to fill gaps in skills.

Writing manuscript

contributors to this section:
Caveat that GitHub has been called out for not being so user-friendly for manuscript development

(Ram 2013). But getting better? Tools that link with GitHub have been developed with synchronous writing in mind. HackMD provides a collaborative writing platform based on Markdown that integrates with GitHub.
We used this platform early on in the process of writing this manuscript to generate an outline.

GitHub in EcoEvo examples (Part 3)

Open science discussion

Contributors to this section: https://github.community/ GitHub new discussion tool

Project continuity

Contributors to this section: 'thus preserving the long-term integrity of the project even as collaborations form and shift.' (Hampton et al. 2015) better to have old versions on GitHub than on somebody's personal hard drive!

Asycnhronous working

Contributors to this section:

GitHub organizations

Contributors to this section: Lab organization as a place to house research compendia as well as codes of conduct, protocols, training documents, etc. (documents that evolve over time and are shaped collaboratively) Students can have full ownership over repositories in organization, but stay with the lab after they've left.

Misc other uses

Contributors to this section: Developing data standards - This is pretty specific case from our team at Berkeley Lab, so I can write about it in a sentence or two here. Code review rOpenSci's code review process, and also caught mistakes in code of published papers that could have been caught in peer code review. Also maybe say something about ReproHack.

Discussion

General pargrah on what GitHub can enable in EcoEvo

General paragraph on GitHub on how, given all the potential uses of GitHub, it can enable more collaborative EcoEvo research Despite all the awesomeness of GitHub, there are still plenty of times when you might look to other plantforms for collaboration

Why aren't more people using GitHub?

Learning to use Github requires time, but the payoff is *[may be?]* worth it. Time vs. effort examples or analyses to demonstrate the payoff can help drive the point home to convince people to learn these tools

Limitations

Our own limitations since we are mostly writing from the EcoEvo perspective/ additional github limitation Reliance on R since we are generally in EcoEvo Discussion of free vs. paid plans. When projects get highly collaborative may have to add / pay for accounts. At this point, little difference between paid and free.

Using GitHub is a good start, but lots of practices to make repo more user friendly

end off with our 5/10 tips for how to gain knowledge/practice with GitHub here

Conclusion

Acknowledgements

This manuscript arose from a hackathon at the Society for Open, Reliable, and Transparent Ecology and Evolution (SORTEE) virtual meeting in 2020.

(add funding as needed!)

Code and data availability

The source code and data for this manuscript are available at https://github.com/SORTEE-Github-Hackathon/manuscript.

References

1. **Build software better, together**
   GitHub
   https://github.com

2. **Excuse Me, Do You Have a Moment to Talk About Version Control?**
   Jennifer Bryan
   The American Statistician (2018-04-24) https://doi.org/gdhzdp
   DOI: 10.1080/00031305.2017.1399928

3. **Git can facilitate greater reproducibility and increased transparency in science**
   Karthik Ram
   Source Code for Biology and Medicine (2013-02-28) https://doi.org/krv
   DOI: 10.1186/1751-0473-8-7 · PMID: 23448176 · PMCID: PMC3639880

4. **Our path to better science in less time using open data science tools**
   Julia SStewart Lowndes, Benjamin D Best, Courtney Scarborough, Jamie C Afflerbach, Melanie R Frazier, Casey C O'Hara, Ning Jiang, Benjamin S Halpern
   Nature Ecology & Evolution (2017-06) https://doi.org/gc4jb3
   DOI: 10.1038/s41559-017-0160 · PMID: 28812630

5. **A Quick Introduction to Version Control with Git and GitHub**
   John D Blischak, Emily R Davenport, Greg Wilson
   PLOS Computational Biology (2016-01-19) https://doi.org/gbqsnf
   DOI: 10.1371/journal.pcbi.1004668 · PMID: 26785377 · PMCID: PMC4718703

6. **Ten Simple Rules for Taking Advantage of Git and GitHub**
   Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe da Veiga Leprevost, Christian Fufezan, Tobias Ternent, Stephen J Eglen, Daniel S Katz, … Juan Antonio Vizcaíno
   PLOS Computational Biology (2016-07-14) https://doi.org/gbrb39
   DOI: 10.1371/journal.pcbi.1004947 · PMID: 27415786 · PMCID: PMC4945047

7. **Let's Git started | Happy Git and GitHub for the useR**
   Jenny Bryan Hester the STAT 545 TAs, Jim
   https://happygitwithr.com/

8. **Democratic databases: science on GitHub**
   Jeffrey Perkel
   Nature (2016-10-06) https://doi.org/gdz6dq
   DOI: 10.1038/538127a · PMID: 27708327

9. **A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats**
   Robert Crystal-Ornelas, Charuleka Varadharajan, Ben Bond-Lamberty, Kristin Boye, Madison Burrus, Shreyas Cholia, Michael Crow, Joan Damerow, Ranjeet Devarakonda, Kim S Ely, … Deborah A Agarwal
   Earth and Space Science (2021-08) https://doi.org/gmbs9c
   DOI: 10.1029/2021ea001797

10. **Barely sufficient practices in scientific computing**
    Graham Lee, Sebastian Bacon, Ian Bush, Laura Fortunato, David Gavaghan, Thibault Lestang, Caroline Morton, Martin Robinson, Philippe Rocca-Serra, Susanna-Assunta Sansone, Helena

Webb
Patterns (2021-02) https://doi.org/gjpcb6
DOI: 10.1016/j.patter.2021.100206 · PMID: 33659915 · PMCID: PMC7892476