**Using GitHub as a Data and Research Hub for Individuals, Small Groups, and Community Initiatives**

Author: Adam Chau  
Date: 10/14/2019

**Background and Introduction**

While primarily used by software engineers and architects, GitHub—at its heart a technical sharing platform—has all of the tools, and a community minded spirit behind it, to be used by individuals and groups to share, publish, and own data in a transparent and distributed way (at full scale).

GitHub, as a core service, is an online software service and set of tools primarily for application developers/engineers, and technical individuals. It is an online service built on Git, which is a “free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.[[1]](#endnote-1)” The service is used by some of the largest, and smallest companies, as well as individuals to manage, store, review, document, and most importantly share software code.

While the toolset could at first sound daunting to those not familiar with software development, the tools and concepts are easy to get, work intuitively, and have already been used outside of software development like publishing patents[[2]](#endnote-2) and allowing mathematicians to collaborate and create a textbook[[3]](#endnote-3).

The intent of this short paper is not to go in-depth into all of the workings of GitHub (however some are discussed), but to look at ways that GitHub can help solve issues in data sharing and ownership especially in de-centralized or scattered groups or those that come out of socially minded programs and initiatives. At the same time it looks to also discuss the way GitHub can level the playing the field for those same groups, and individuals, in publishing research, data, theory, and other ideas.

**What is GitHub?**

There are already articles and posts on GitHub, especially directing non-technical people in understanding the key components of GitHub[[4]](#endnote-4). At a basic level, someone should understand these core key concepts:

1. *Repository*: A repository is the main area (like a folder on a computer system) which houses all the documents for a project. These can be all types of documents including word processing documents, text files, images, and database files to name a few. A repository keeps state of all the documents and versions of those documents, who edited them, as well as having a workflow already instituted for group collaboration and managing the collaboration like adding to an existing work. A repository can be “forked” so that it could be used by another group or individual, where they can then use the repository as a base for their own new project. If made public, repositories can be downloaded by anyone with full transparency to changes and data.
2. *Collaborators*: Individuals who can edit and add to a repository.
3. *Branch*: A branch, like its name, is a new branch of a repository, which can be edited and changed, with those changes being merged back into the main repository branch (often called the “Master” branch) with approval. This allows multiple collaborators to be able to work on the different features or portions of a project.

**The Problem with Data Sharing and Ownership**

In any organization there exists policies and procedures to work with data: where it is stored, who can create data, who has access to read it, and who has access to update it. For smaller organizations, teams, and groups, this can sometimes present issues because of resources and infrastructure, including the following:

1. *Data Loss*: When a small group of people create an organization and then create a website, host information for the community, but then after some time the organization disbands, and the website and associated information is lost forever, even though it contained valuable information for the community.
2. *Data Ownership*:At times, there exists a relationship between organizations and individuals, where both sides have a vested interest in the same data, or type of data, however, there exists disagreements on who controls and owns that data. Even if an organization makes information available to the public, it’s still in the domain of the organization, versus the community and individuals.
3. *Data Transparency:* Simply because so much data is in the hands of one organization, or a small group of people, data transparency—who’s modified the data, when was it modified—can be hard, if not impossible to trace.

**How GitHub Can Remove Barriers So That Data Can Be Shared Better Among the Community with More Equitable Data Ownership**

Because GitHub is a collaborative space where data can truly be owned by everyone, along with the tools needed to manage that data, it provides an ideal space for individuals, groups, and organizations to utilize for community minded projects, removing barriers to some of the problems listed earlier (data loss, ownership of the data, and data transparency).

To better illustrate, below are a few examples of how groups could make use of GitHub for community minded projects:

*Example 1:*

A group of four international adoptees are doing research on a group of adoptees from country X and how many are located on the East Coast, and breakdowns of those adoptees by age, educational level reached, and field of employment. Once they have done their research and created their findings they:

1. Publish their findings on GitHub in a new repository including some documents and spreadsheets. They make the repository public so that anyone can download and own all documents and versioning history.
2. A group on the West Coast hears about their research and wants to create their own repository except on West Coast adoptees from Country X. They fork the repository for use in their own project making use of some of the templates and information from the East Coast group.

In this example, the group on the West Coast was able to easily utilize the East Coast group’s templates by forking their repository as a base for their own project.

*Example 2*

A group focused on LGBTQ rights that exists in the Midwest has decided to create a list of all LGBTQ run restaurants in their respective states. They decide that there should be no one person in charge of the list but that the group itself should be the keepers of the list. After publishing their spreadsheet on GtiHub and making the members from the group Collaborators:

1. All group members cloned and synced the repository.
2. Group member A created a branch in the repository called “Member A Branch” and Group member B created a branch called “Member B Branch”. Both found two new restaurants to add to the list, added it to their branch, and then submitted a request to have those added into the Master branch. Both requests were approved by other members, and the additions were added.
3. A few months later, Group member C saw that one of the restaurants closed, so they updated their branch with the deleted entry, and submitted the request to be approved into the Master branch. After approval (by member D) member A saw that one of the listings they submitted was no longer there, but were able to go into the version history, see the change, and why it happened. If someone decided later that the deletion was a mistake, it could be easily rolled back and updated.
4. A year after the project was rolled out, one of the members who helped start it, and create the original GitHub project decided to leave the project, and no longer wanted to have the repository under their account. The member was able to easily transfer the repository to another member’s account with all history and information being intact.  
   1. In the worst case scenario, if a repo is set as public, it can be cloned and then used to create a completely new repository with the same data, but without some data like issues, pull requests, and other GitHub elements (however copies over the Git data).

**A Platform to Publish**

Simply because GitHub exists in part to help with the advancement of sharing data by making that data public, where it can be shared *and* scrutinized, it can be used as a platform to those groups and individuals who may not work or exist within academic or scientific institutions, to publish their data, research, and theories, within an environment which can lend itself to having more validity (at least in appearance). While this can be subjective, and cannot replace good research, critical thought, etc., and doesn’t necessarily invalidate other means of self-publication (say for instance a blog or website), the mere fact that a repository can be made public, shared, and added to (if allowed)—should be an impetus for publishing sound data and research, regardless of the topic.

At the same time, GitHub can be used to publish theories and data for topics which do not have a place in traditional academic or community minded research, or where the publishing of those theories, data, and research is gate-kept in a way that excludes the majority of individuals, even if they may have compelling theories and data to add to existing research and discussion.

**Conclusion**

For individuals, small and non-profit groups, and community minded initiatives and organizations, using GitHub as a way to publish, manage, and share data, while allowing for the de-centralization of that data, can be an incredible tool to push the advancement of research and community involvement.

With more advances in the technology and workflow, as well as the perception of what, and where valid research and thoughts live, GitHub can only help improve open source and community minded projects, as well as be used for a platform for more individual research, theories, and papers of a more academic nature, regardless of the topic.

1. https://git-scm.com/ [↑](#endnote-ref-1)
2. https://readwrite.com/2013/11/08/seven-ways-to-use-github-that-arent-coding/ [↑](#endnote-ref-2)
3. https://github.com/HoTT/book [↑](#endnote-ref-3)
4. https://medium.com/crowdbotics/a-dead-simple-intro-to-github-for-the-non-technical-f9d56410a856 [↑](#endnote-ref-4)