`commit` -ment issues with Git: investigating & archiving y’alls work

Sarah Nguyen, Vicky Steeves (Presenting) & Genevieve Milliken

csv,conf,v5 | 2020.05.14

Slides: osf.io/uftkn

#IASGE

About IASGE

(ice-age)

#IASGE

Who are we?

New York University, Division of Libraries!

In addition to the core team (below), this project wouldn't be possible without the time and efforts of colleagues in Digital Library Technology Services

Vicky Steeves
Project Lead

Genevieve Milliken
Research Scientist

Sarah Nguyen
Research Scientist

#IASGE

Project overview

An Alfred P. Sloan Foundation funded project, IASGE (pronounced `ice-age`) has two main streams of work:

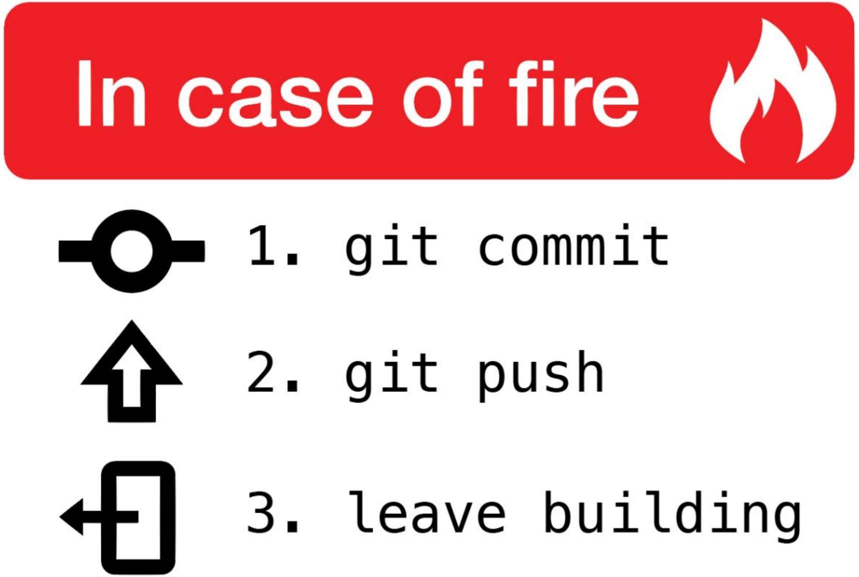
1) Study how academics/folks in academia are using Git and Git hosting platforms and how these tools could be better aligned with their needs
2) Evaluate the extent to which the scholarship on Git hosting platforms is being preserved by professionals, and write an archival spec

The results of this project aim to inform the way code and annotations on Git hosting platforms move from a phase where they are highly active and collaborative, to a state where they are stable, permanently citable, and under **active, professional preservation**.

What is Git?

- Git is a revision control system, a program to manage your source code history
- It is strictly a command-line tool
- Revision control systems let us compare, restore, and merge changes to our [plain-text] stuff

This is hugely important for collaboration and transparency!



In case of fire

1. git commit

2. git push

3. leave building

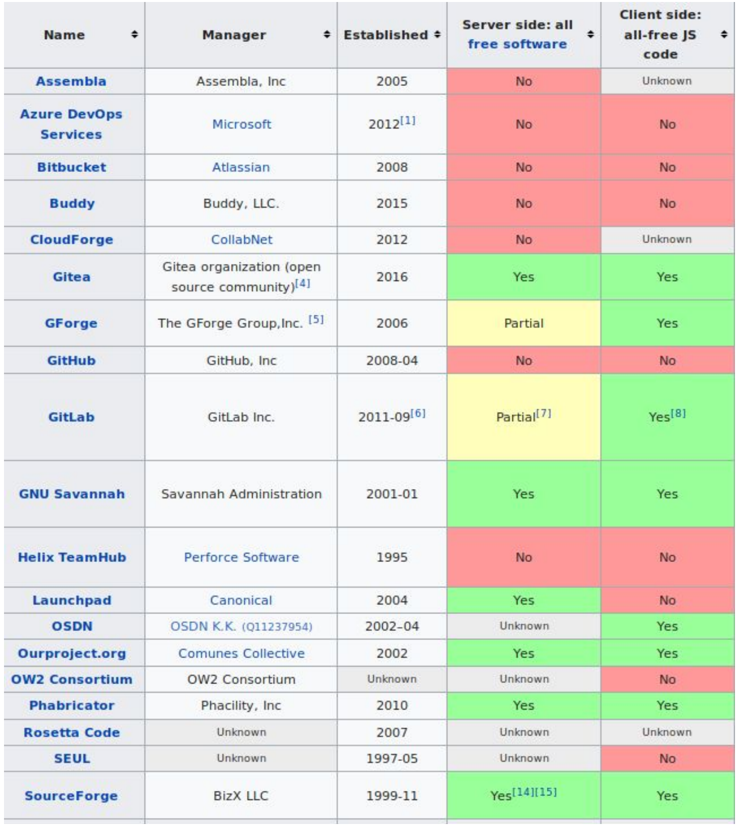
What are “Git Hosting Platforms”

Literally, places that host git repositories on the web

They are NOT the same as Git, but rather are places where you can upload Git repositories with some additional features

The most popular include:

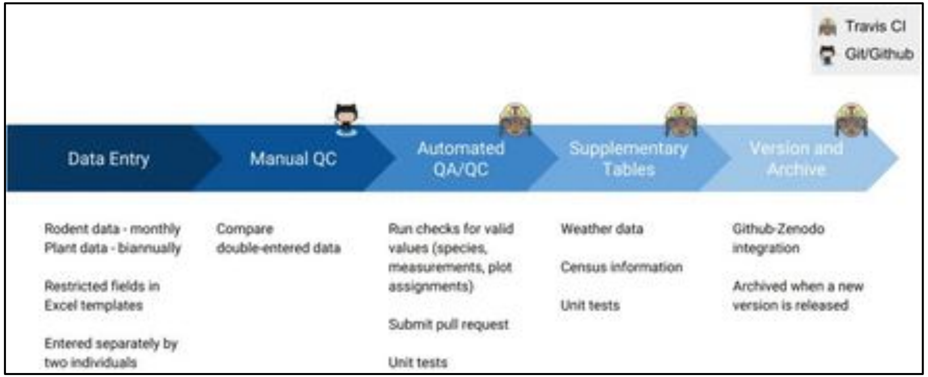
1. GitHub
2. GitLab
3. Bitbucket
4. Sourceforge



Name	Manager	Established	Server side: all free software	Client side: all-free JS code
Assembla	Assembla, Inc	2005	No	Unknown
Azure DevOps Services	Microsoft	2012[1]	No	No
Bitbucket	Atlassian	2008	No	No
Buddy	Buddy, LLC.	2015	No	No
CloudForge	CollabNet	2012	No	Unknown
Gitea	Gitea organization (open source community)[4]	2016	Yes	Yes
GForge	The GForge Group,Inc. [5]	2006	Partial	Yes
GitHub	GitHub, Inc	2008-04	No	No
GitLab	GitLab Inc.	2011-09[6]	Partial[7]	Yes[8]
GNU Savannah	Savannah Administration	2001-01	Yes	Yes
Helix TeamHub	Perforce Software	1995	No	No
Launchpad	Canonical	2004	Yes	No
OSDN	OSDN K.K. (Q11237954)	2002–04	Unknown	Yes
Ourproject.org	Comunes Collective	2002	Yes	Yes
OW2 Consortium	OW2 Consortium	Unknown	Unknown	No
Phabricator	Phacility, Inc	2010	Yes	Yes
Rosetta Code	Unknown	2007	Unknown	Unknown
SEUL	Unknown	1997-05	Unknown	No
SourceForge	BizX LLC	1999-11	Yes[14][15]	Yes

Examples of “scholarly Git” usage

1. Publishing code and data as supplementary materials
2. Quality assurance workflows for data analysis
3. Journal infrastructure with peer review



Travis CI
Git/Github

Data Entry
- Rodent data - monthly
  Plant data - biannually
- Restricted fields in Excel templates
- Entered separately by two individuals

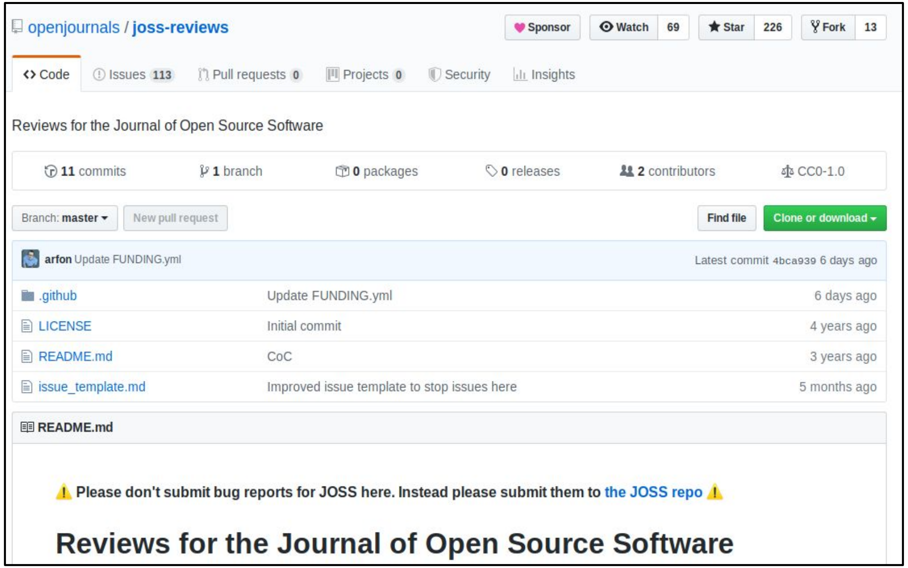
Manual QC
- Compare double-entered data

Automated QA/QC
- Run checks for valid values (species, measurements, plot assignments)
- Submit pull request
- Unit tests

Supplementary Tables
- Weather data
- Census information
- Unit tests

Version and Archive
- Github-Zenodo integration
- Archived when a new version is released

https://www.biorxiv.org/content/10.1101/344804v2.full



openjournals / joss-reviews

Reviews for the Journal of Open Source Software

arfon Update FUNDING.yml — Latest commit 4bca939 6 days ago

.github	Update FUNDING.yml	6 days ago
LICENSE	Initial commit	4 years ago
README.md	CoC	3 years ago
issue_template.md	Improved issue template to stop issues here	5 months ago

README.md

⚠️ **Please don't submit bug reports for JOSS here. Instead please submit them to the JOSS repo** ⚠️

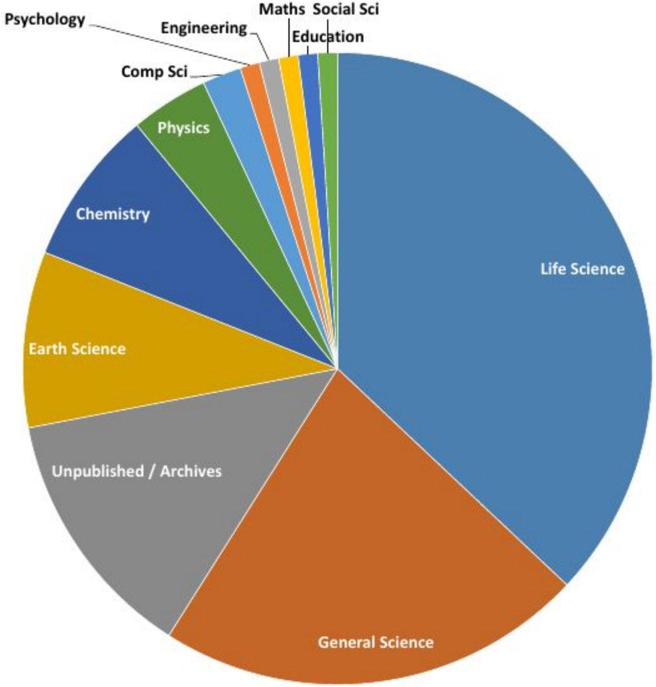
Reviews for the Journal of Open Source Software

#IASGE

Estimated scope of scholarship in GHPs

“Over 5,000 Github software repositories have been identified as research software according to the criteria explained previously: either a research publication referenced the software repository, or the software repository referenced a research publication.”

-- Hasselbring, Wilhelm, et al. “FAIR and Open Computer Science Research Software.” ArXiv:1908.05986 [Cs], Aug. 2019. http://arxiv.org/abs/1908.05986.

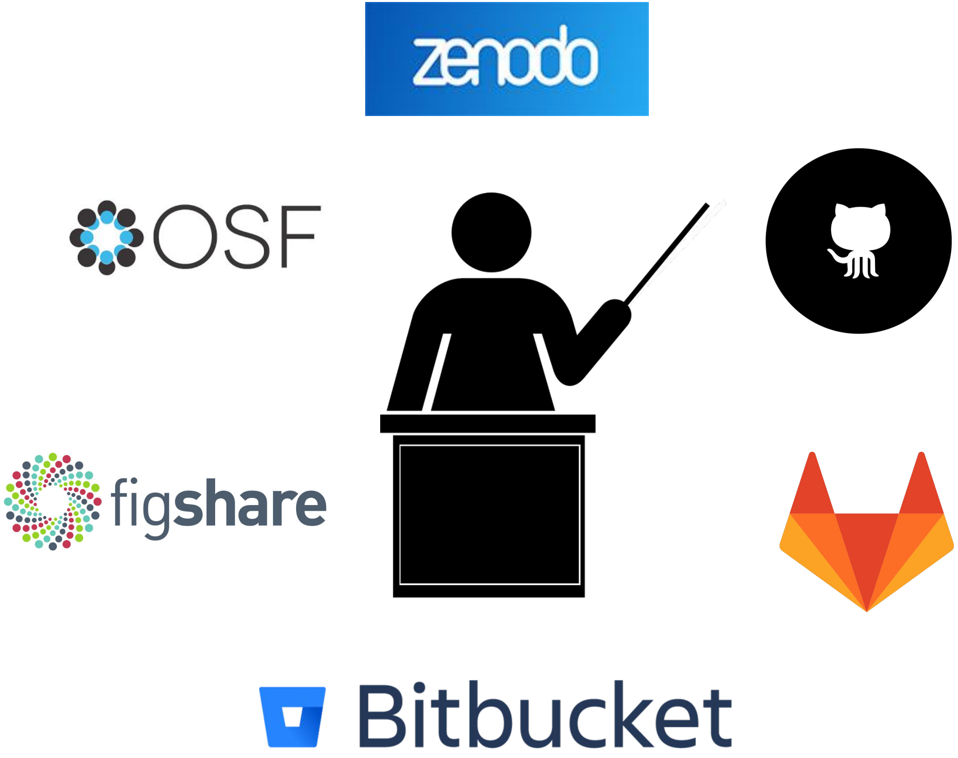


Research areas of publications cited from Github repositories from arXiv: 1909.05986

#IASGE

What’s the problem we want to solve

- Researchers use a variety of scholarly tools on the web during the research process, which includes designing, developing, and refining (through versioning) source code
- This source code is contextualized by the “scholarly ephemera” associated with it (e.g. issue disc.)
- No project currently captures both source code and scholarly ephemera

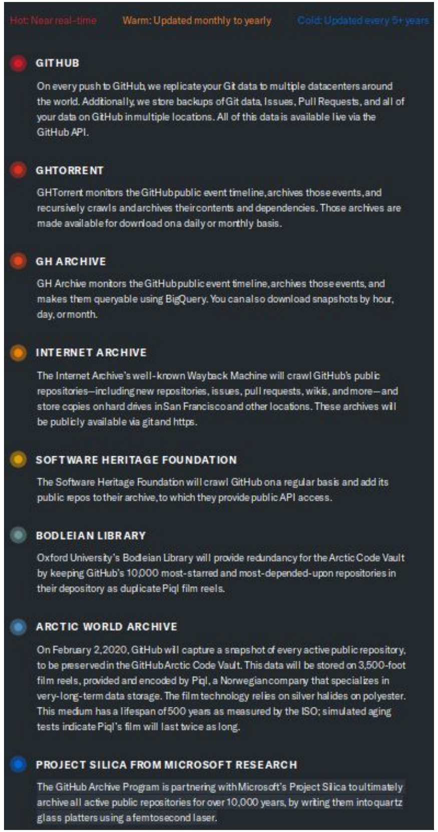


🐘 in the room: GitHub’s Archiving Program

There are a lot of open questions about GitHub’s Archive Program which are probably shared by people in this room

The fact remains that none of the partners or solutions here capture the ephemera + source code reliably together, which we posit as important for usability in the long-term

Also, read DSHR’s takedown: blog.dshr.org/2019/11/seeds-or-code.html



Hot: Near real-time Warm: Updated monthly to yearly Cold: Updated every 5+ years

● **GITHUB**

On every push to GitHub, we replicate your Git data to multiple datacenters around the world. Additionally, we store backups of Git data, Issues, Pull Requests, and all of your data on GitHub in multiple locations. All of this data is available live via the GitHub API.

● **GHTORRENT**

GHTorrent monitors the GitHub public event timeline, archives those events, and recursively crawls and archives their contents and dependencies. Those archives are made available for download on a daily or monthly basis.

● **GH ARCHIVE**

GH Archive monitors the GitHub public event timeline, archives those events, and makes them queryable using BigQuery. You can also download snapshots by hour, day, or month.

● **INTERNET ARCHIVE**

The Internet Archive’s well-known Wayback Machine will crawl GitHub’s public repositories—including new repositories, issues, pull requests, wikis, and more—and store copies on hard drives in San Francisco and other locations. These archives will be publicly available via git and https.

● **SOFTWARE HERITAGE FOUNDATION**

The Software Heritage Foundation will crawl GitHub on a regular basis and add its public repos to their archive, to which they provide public API access.

● **BODLEIAN LIBRARY**

Oxford University’s Bodleian Library will provide redundancy for the Arctic Code Vault by keeping GitHub’s 10,000 most-starred and most-depended-upon repositories in their depository as duplicate Piql film reels.

● **ARCTIC WORLD ARCHIVE**

On February 2, 2020, GitHub will capture a snapshot of every active public repository, to be preserved in the GitHub Arctic Code Vault. This data will be stored on 3,500-foot film reels, provided and encoded by Piql, a Norwegian company that specializes in very-long-term data storage. The film technology relies on silver halides on polyester. This medium has a lifespan of 500 years as measured by the ISO; simulated aging tests indicate Piql’s film will last twice as long.

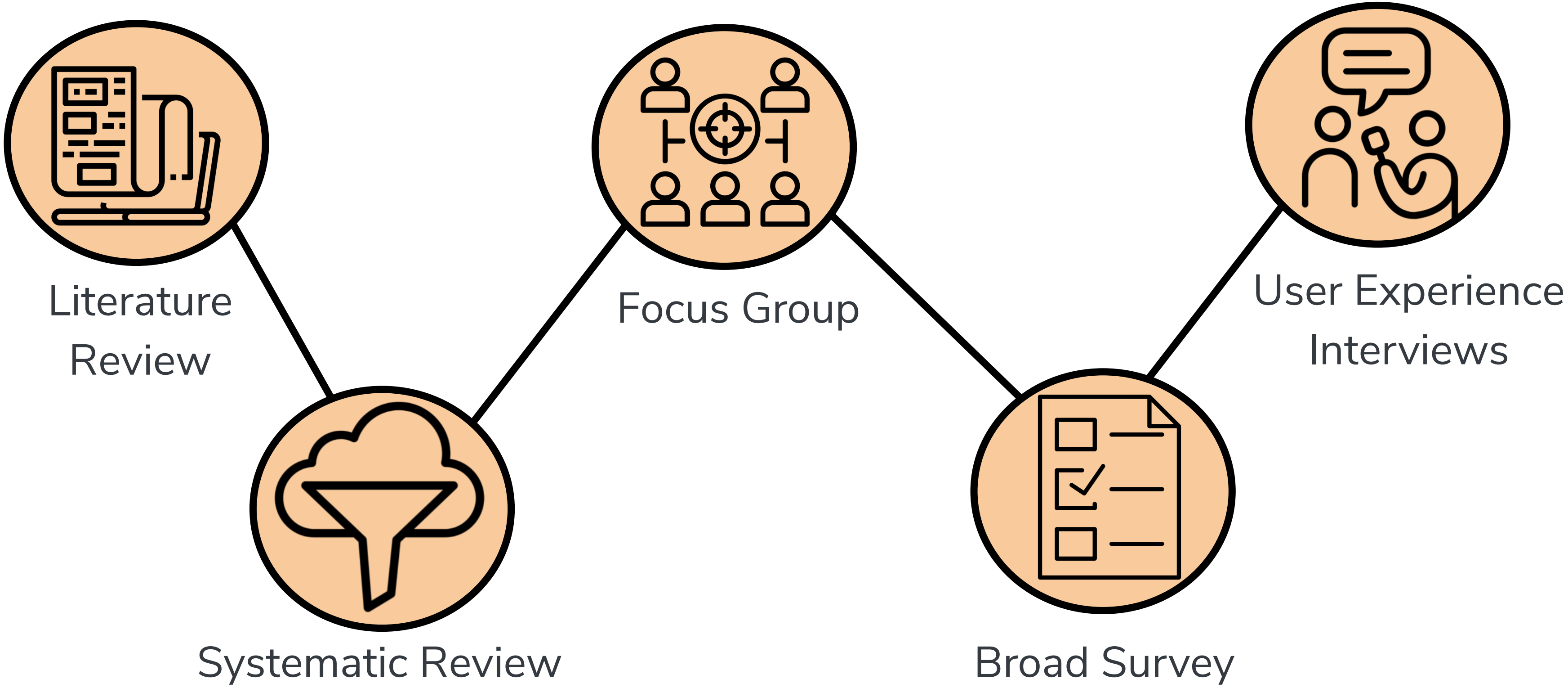
● **PROJECT SILICA FROM MICROSOFT RESEARCH**

The GitHub Archive Program is partnering with Microsoft’s Project Silica to ultimately archive all active public repositories for over 10,000 years, by writing them into quartz glass platters using a femtosecond laser.

Gap Analysis to Understand Scholars Using Git

#IASGE

Researching the Scholarly Git Experience



Literature Review

Systematic Review

Focus Group

Broad Survey

User Experience Interviews

Research Questions

How are scholars currently using this toolkit of Git + GHPs?

How can features in Git + GHPs serve scholarship?

How can we improve teaching Git for minimal users?

Follow our IASGE blog for updates

investigating-archiving-git.gitlab.io/updates/

Examples of scholarly Git* usage

Git Experiences	Related GHP Feature	Related Git Commands

Version control	- Commit logs
	- Branches
		`git log`
		`git diff`

Community & collaboration	- Issue Tracker
	- Pull requests
		`git add <files>`
		`git commit -m "[message]"`
		`git push`

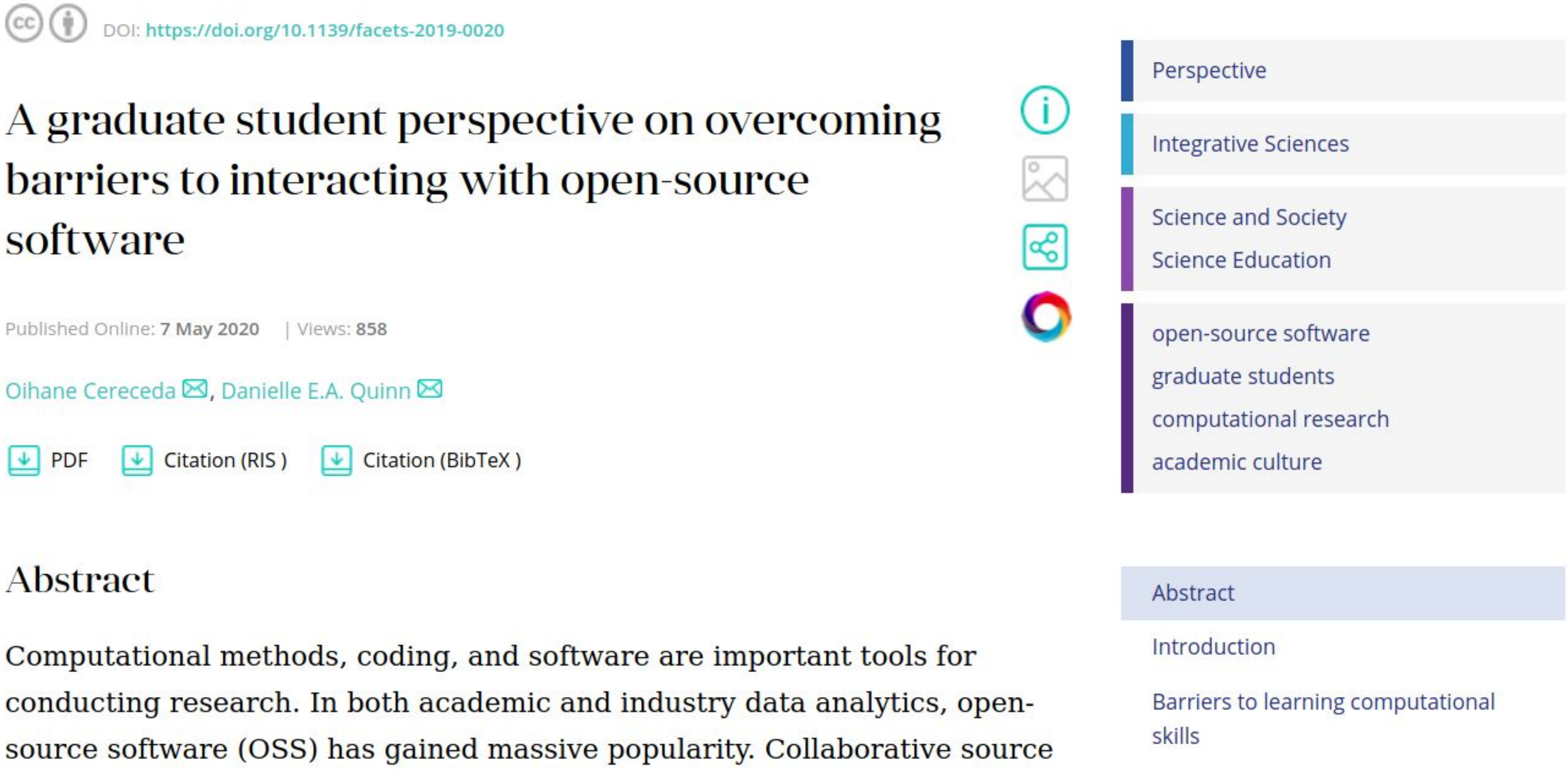
Method tracking	- README - Wiki - Posts - Commit logs	`git commit logs`

Education	- README
	- Wiki
	- Issue Tracker
	- Pull requests
		`open-issues`
		`close-issues`
		`list-issues`
		`check-review`

Data processing	- Continuous integration	(various)

Reproducibility	- README
	- Continuous integration
		`git clone`
		`git pull`

Publishing	- Pages services - README	(various)



DOI: https://doi.org/10.1139/facets-2019-0020

A graduate student perspective on overcoming barriers to interacting with open-source software

Published Online: 7 May 2020 | Views: 858

Oihane Cereceda ✉, Danielle E.A. Quinn ✉

Perspective

Integrative Sciences

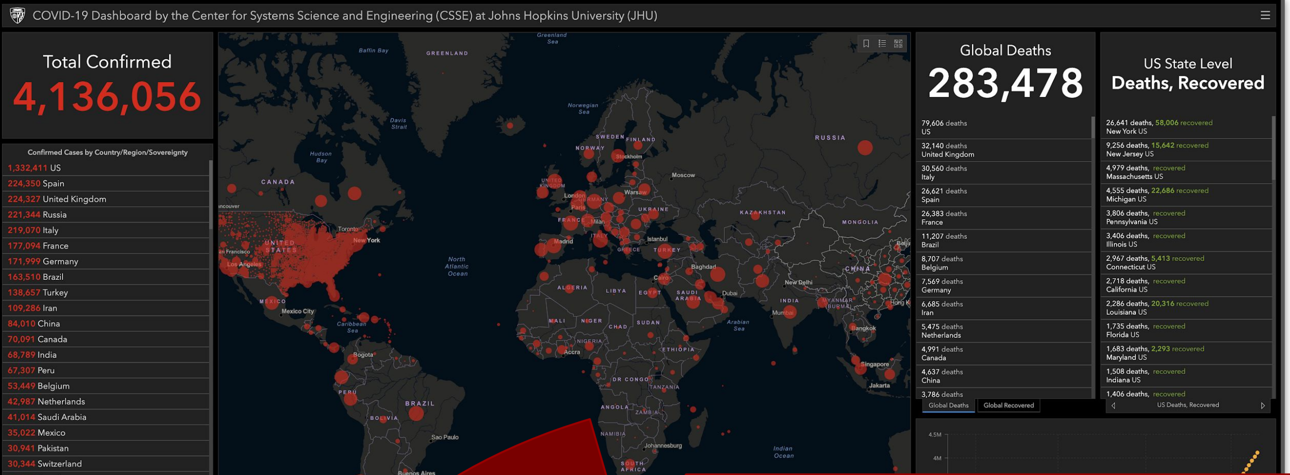
Science and Society
Science Education

open-source software
graduate students
computational research
academic culture

Abstract

Computational methods, coding, and software are important tools for conducting research. In both academic and industry data analytics, open-source software (OSS) has gained massive popularity. Collaborative source

Many academic institutions are using GitHub to share COVID-19 Data



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Total Confirmed

4,136,056

Confirmed Cases by Country/Region/Sovereignty

1,332,411	US
224,350	Spain
224,327	United Kingdom
221,344	Russia
219,070	Italy
177,094	France
171,999	Germany
163,510	Brazil
138,657	Turkey
109,286	Iran
84,010	China
70,091	Canada
68,789	India
67,307	Peru
53,449	Belgium
42,987	Netherlands
41,014	Saudi Arabia
35,022	Mexico
30,941	Pakistan
30,344	Switzerland
29,559	Ecuador
28,866	Chile
27,679	Portugal
26,670	Sweden
23,906	Belarus
23,822	Singapore
23,623	Qatar
22,996	Ireland
18,198	United Arab Emirates

Admin0 Admin1 Admin2

Last Updated at (M/D/YYYY)
5/11/2020, 10:32:31 AM

Cumulative Confirmed Cases | Ac | ase-Fatality Ratio | Testing Rate | Hospitalization Rate

187
countries/regions

Mobile Version: Here.
mation Support: Esri Living Atlas team and JHU APL. Contact US. FAQ.

, CDC, ECDC, NHC, DXY, 1point3acres, Worldometers.info, BNO, the COVID Tracking Project
d local media reports. Read more in this blog.

le database: GitHub: Here. Feature layer: Here.

Global Deaths

283,478

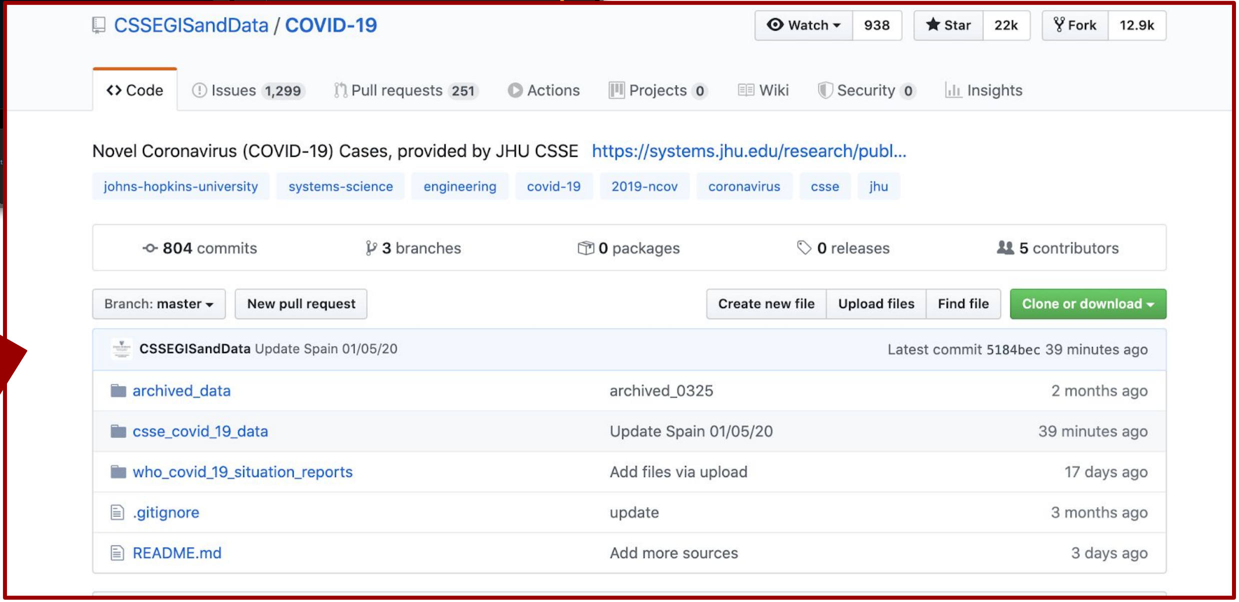
79,606 deaths	US
32,140 deaths	United Kingdom
30,560 deaths	Italy
26,621 deaths	Spain
26,383 deaths	France
11,207 deaths	Brazil
8,707 deaths	Belgium
7,569 deaths	Germany
6,685 deaths	Iran
5,475 deaths	Netherlands
4,991 deaths	Canada
4,637 deaths	China
3,786 deaths	

Global Deaths Global Recovered

US State Level
Deaths, Recovered

26,641 deaths, 58,006 recovered	New York US
9,256 deaths, 15,642 recovered	New Jersey US
4,979 deaths, recovered	Massachusetts US
4,555 deaths, 22,686 recovered	Michigan US
3,806 deaths, recovered	Pennsylvania US
3,406 deaths, recovered	Illinois US
2,967 deaths, 5,413 recovered	Connecticut US
2,718 deaths, recovered	California US
2,286 deaths, 20,316 recovered	Louisiana US
1,735 deaths, recovered	Florida US
1,683 deaths, 2,293 recovered	Maryland US
1,508 deaths, recovered	Indiana US
1,406 deaths, recovered	

US Deaths, Recovered



CSSEGISandData / COVID-19

Watch 938 Star 22k Fork 12.9k

Code Issues 1,299 Pull requests 251 Actions Projects 0 Wiki Security 0 Insights

Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE https://systems.jhu.edu/research/publ...

johns-hopkins-university systems-science engineering covid-19 2019-ncov coronavirus csse jhu

804 commits 3 branches 0 packages 0 releases 5 contributors

CSSEGISandData Update Spain 01/05/20 Latest commit 5184bec 39 minutes ago

archived_data	archived_0325	2 months ago
csse_covid_19_data	Update Spain 01/05/20	39 minutes ago
who_covid_19_situation_reports	Add files via upload	17 days ago
.gitignore	update	3 months ago
README.md	Add more sources	3 days ago

Johns Hopkins University’s Center for Systems Science & Engineering

Academic institutions are using GitHub to share COVID-19 models

16 March 2020 *Imperial College COVID-19 Response Team*

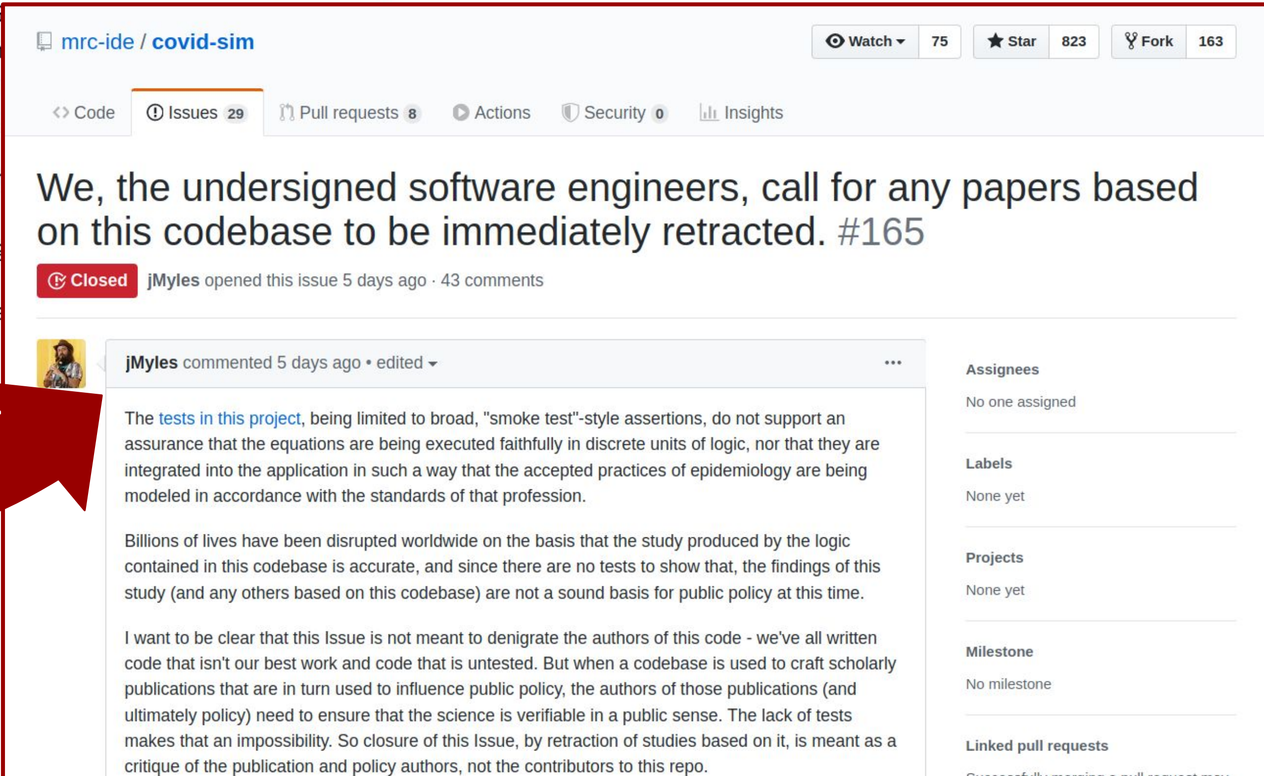
Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand

Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin,
Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, G
Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Ham
Elsland, Hayley Thompson, Robert Verity, k Volz, Haowei
Caroline Walters, Peter Winskill, Char aker, Christl

On behalf of the Imperial College Response Tear

WHO Collaborating Centre fo us Disease Modelling
MRC Centre for Global Infec ease Analysis
Abdul Latif Jameel Institute ease and Emergency Ana
Imperial College London

Correspondence: neil.fergus @imperial.ac.uk



We, the undersigned software engineers, call for any papers based on this codebase to be immediately retracted. #165

jMyles opened this issue 5 days ago · 43 comments

jMyles commented 5 days ago • edited

The tests in this project, being limited to broad, "smoke test"-style assertions, do not support an assurance that the equations are being executed faithfully in discrete units of logic, nor that they are integrated into the application in such a way that the accepted practices of epidemiology are being modeled in accordance with the standards of that profession.

Billions of lives have been disrupted worldwide on the basis that the study produced by the logic contained in this codebase is accurate, and since there are no tests to show that, the findings of this study (and any others based on this codebase) are not a sound basis for public policy at this time.

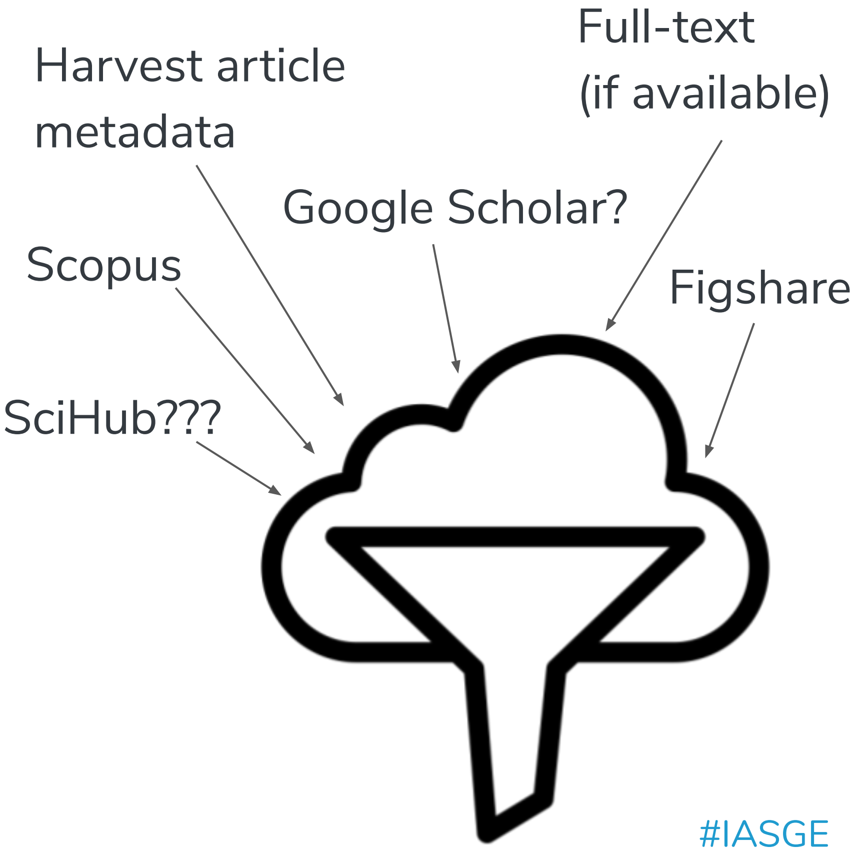
I want to be clear that this Issue is not meant to denigrate the authors of this code - we've all written code that isn't our best work and code that is untested. But when a codebase is used to craft scholarly publications that are in turn used to influence public policy, the authors of those publications (and ultimately policy) need to ensure that the science is verifiable in a public sense. The lack of tests makes that an impossibility. So closure of this Issue, by retraction of studies based on it, is meant as a critique of the publication and policy authors, not the contributors to this repo.

Imperial College of London's MRC Centre for Global Infectious Disease Analysis (MRC GIDA)

Part I: Systematic Review (Quant)

An approach to understand the landscape of published scholarly articles that reference Git repositories.

“All our source code is available on [GitLab], to allow community to reproduce our results, from the training of the networks, until the statistical analyses.” (Perez, 2019)



Full-text (if available)

Harvest article metadata

Google Scholar?

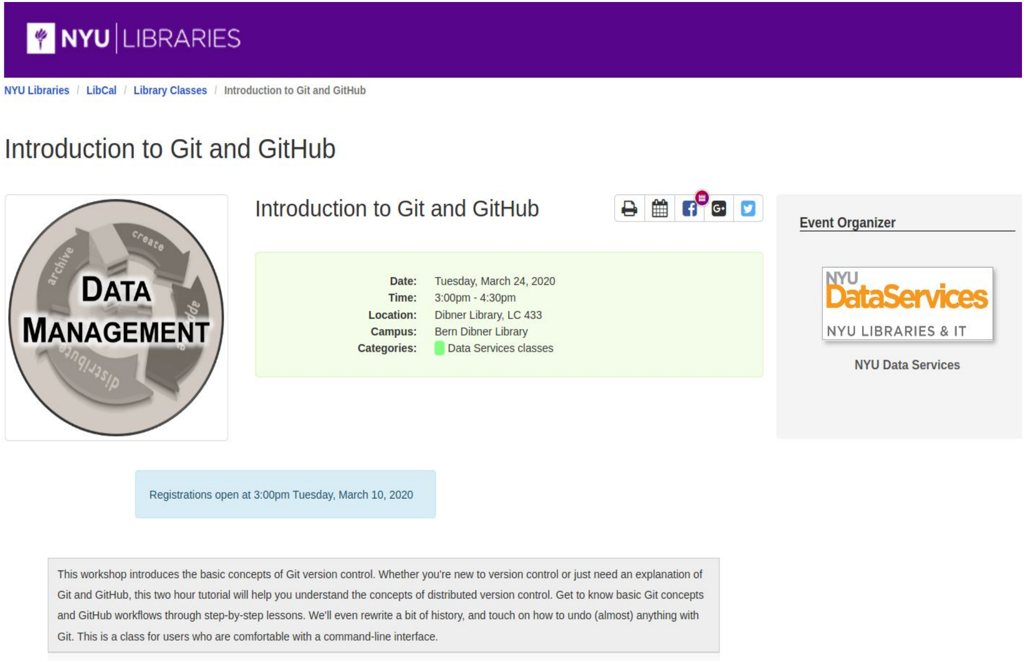
Scopus

Figshare

SciHub???

Part II: Focus Group (Qual)

. . . teachers were surprised by how overwhelming the student enthusiasm is for adopting VCS, [but] they also discovered that they lacked understanding about the system or having confidence in their ability to use it effectively beyond the course.



NYU Libraries / LibCal / Library Classes / Introduction to Git and GitHub

Introduction to Git and GitHub

Introduction to Git and GitHub

Date: Tuesday, March 24, 2020
Time: 3:00pm - 4:30pm
Location: Dibner Library, LC 433
Campus: Bern Dibner Library
Categories: 🟩 Data Services classes

Event Organizer

NYU Data Services

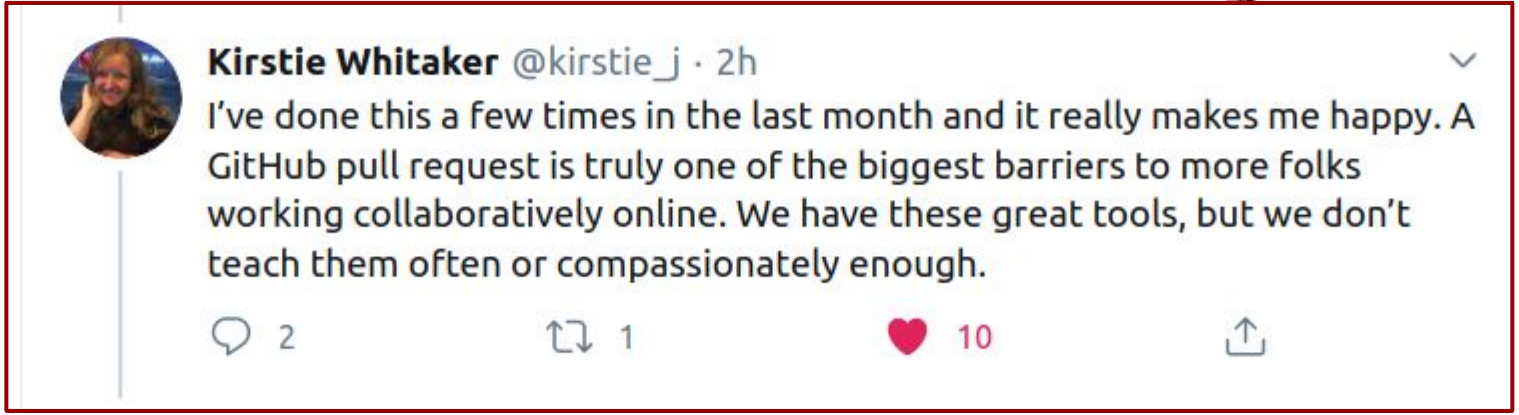
Registrations open at 3:00pm Tuesday, March 10, 2020

This workshop introduces the basic concepts of Git version control. Whether you're new to version control or just need an explanation of Git and GitHub, this two hour tutorial will help you understand the concepts of distributed version control. Get to know basic Git concepts and GitHub workflows through step-by-step lessons. We'll even rewrite a bit of history, and touch on how to undo (almost) anything with Git. This is a class for users who are comfortable with a command-line interface.

——Glassey, R. (2019). Adopting Git/Github Within Teaching: A Survey of Tool Support. *Proceedings of the ACM Conference on Global Computing Education*, 143–149. https://doi.org/10.1145/3300115.3309518

#IASGE

PULL REQUEST
CLONE
?
APPROVED
Scriberia


Kirstie Whitaker @kirstie_j · 2h

I've done this a few times in the last month and it really makes me happy. A GitHub pull request is truly one of the biggest barriers to more folks working collaboratively online. We have these great tools, but we don't teach them often or compassionately enough.

Part III: Survey (Quant)

Target population:
Scholars who use Git across all disciplines & statuses

Goal:
To gather a wide-ranging & comparable census

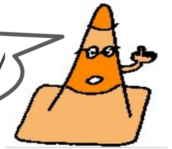
Please participate & share widely!
bit.ly/3aO0ykQ

Closes June 22nd

Themes

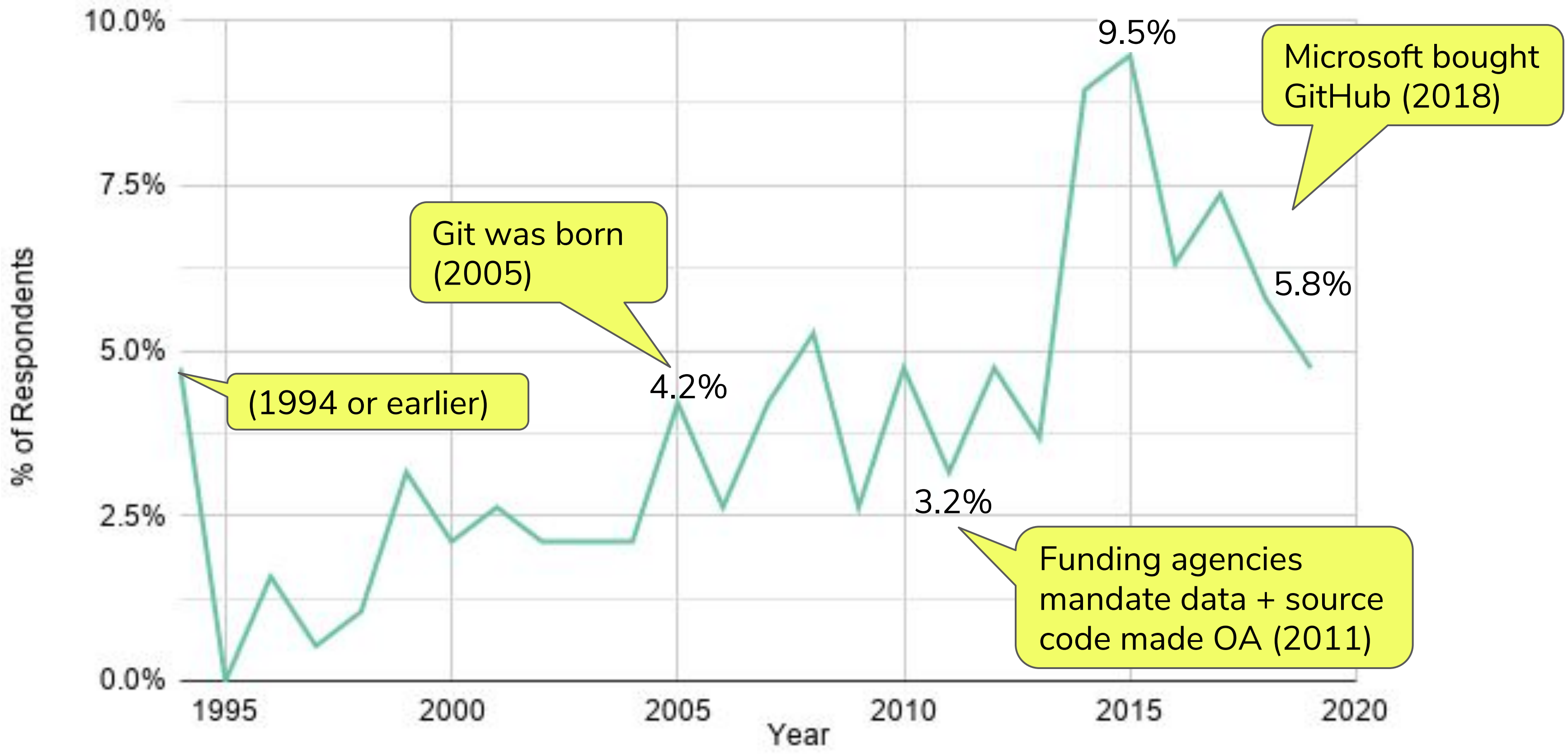
- ❑ Learning Git
- ❑ Teaching Others
- ❑ Daily Use
- ❑ Features on hosting platforms
- ❑ Scholarship
- ❑ Follow-up

Survey - Preliminary Findings



Please participate & share widely
bit.ly/3aO0ykQ

Q: When did you first start using a version control system?



(1994 or earlier)

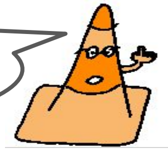
Git was born (2005)

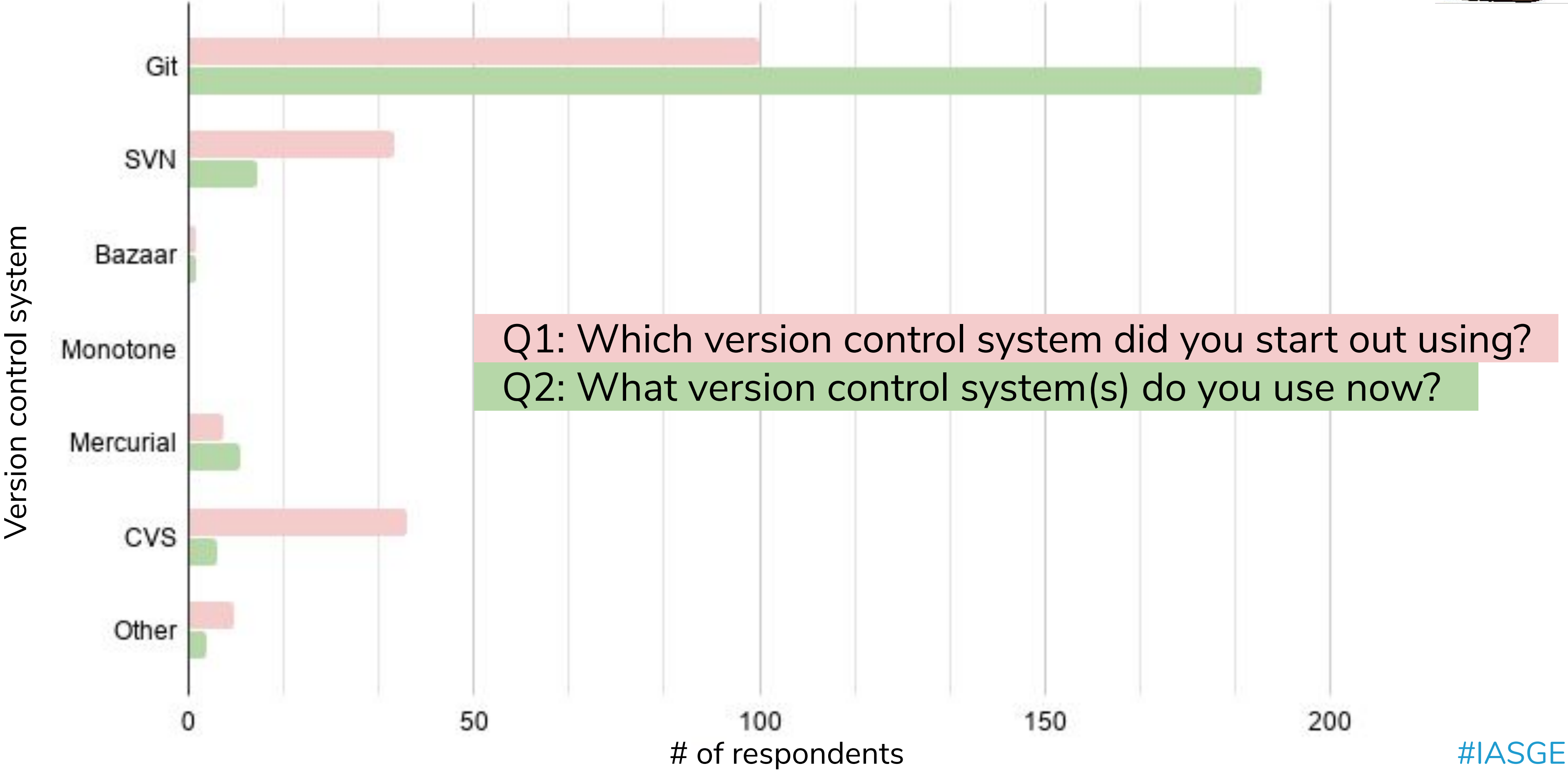
Funding agencies mandate data + source code made OA (2011)

Microsoft bought GitHub (2018)

Survey - Preliminary Findings

Please participate & share widely!
[bit.ly/3aO0ykQ](bit.ly/3aO0ykQ)





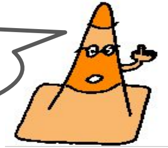
Q1: Which version control system did you start out using?

Q2: What version control system(s) do you use now?

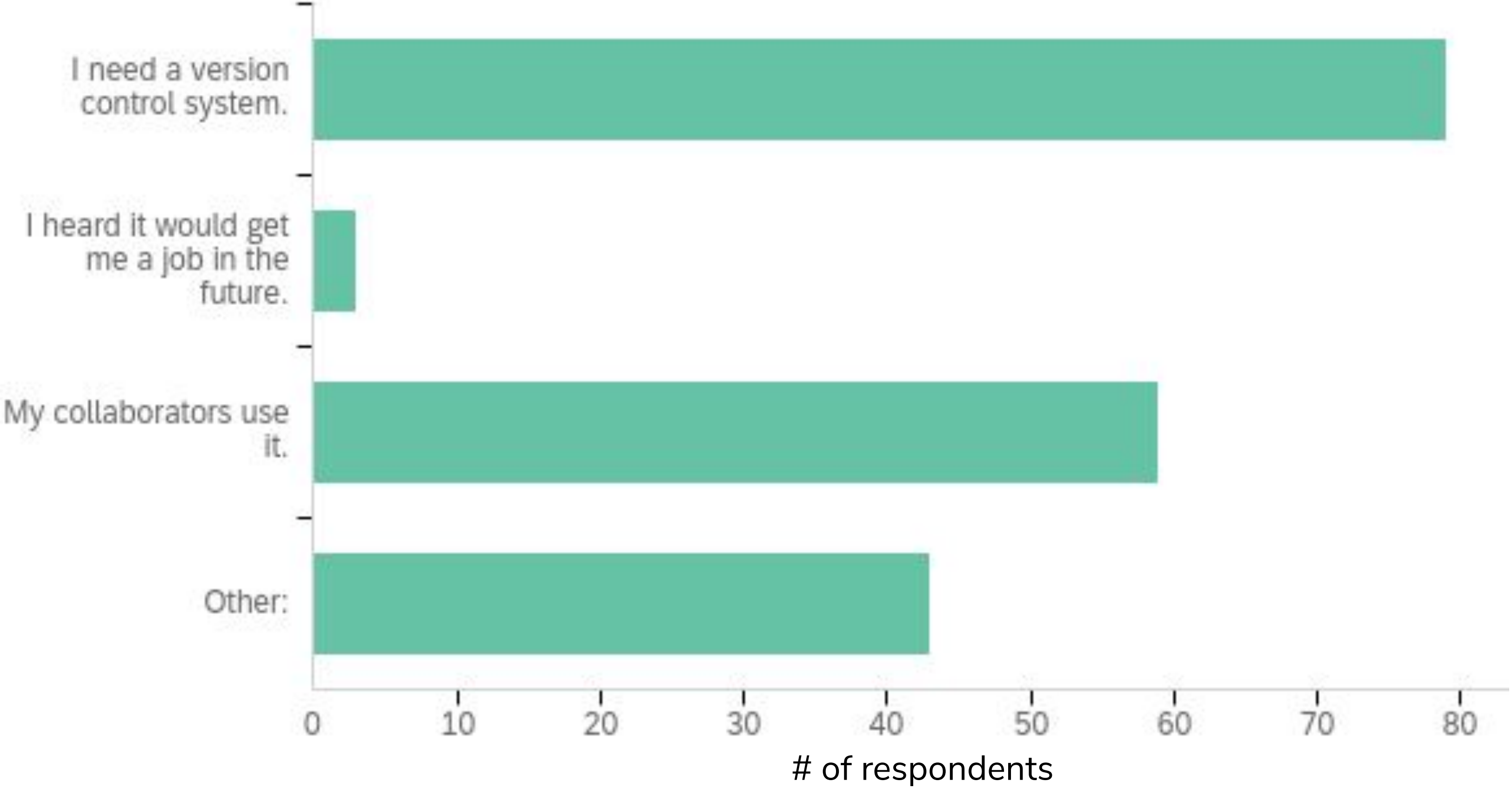
#IASGE

Survey - Preliminary Findings

Please participate & share widely!
bit.ly/3aO0ykQ



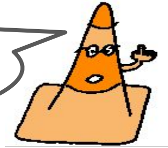
Q: Why did you first enter the world of git and version control?



#IASGE

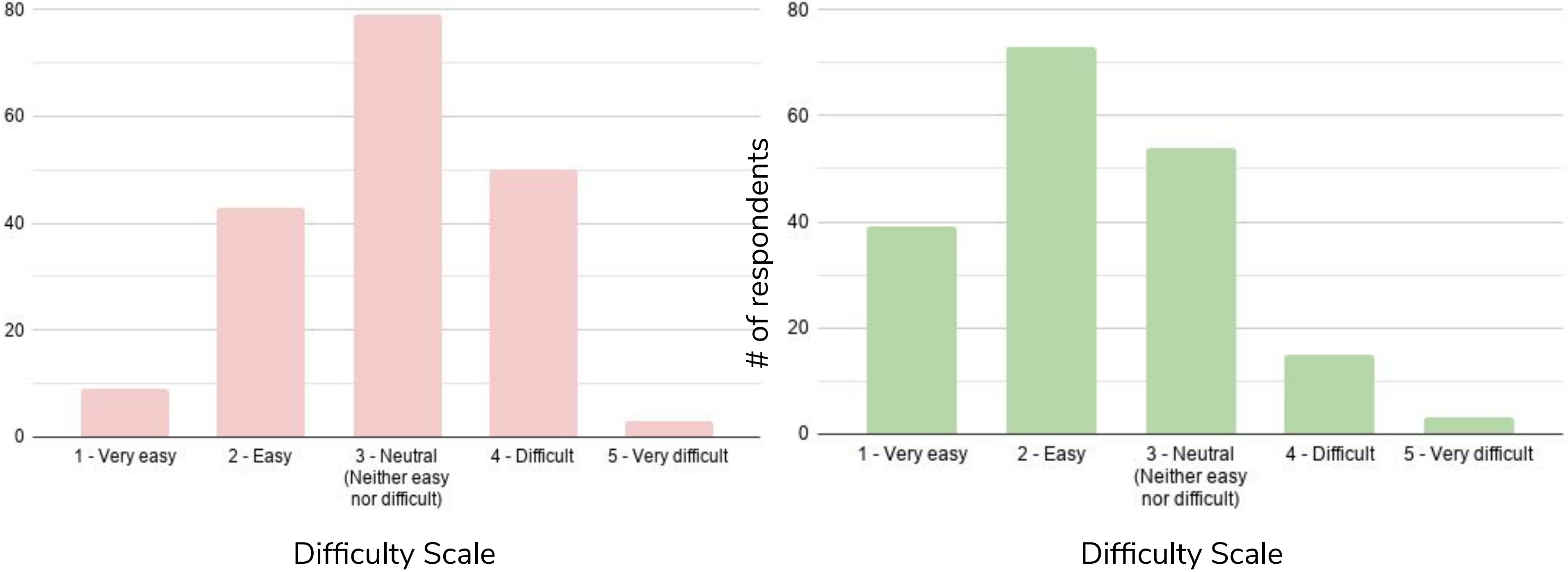
Survey - Preliminary Findings

Please participate & share widely!
bit.ly/3aO0ykQ



Q1: How difficult was it for you to learn how to use git on your local computer?

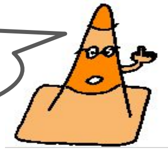
Q2: How difficult was it for you to learn how to use the git hosting platform?

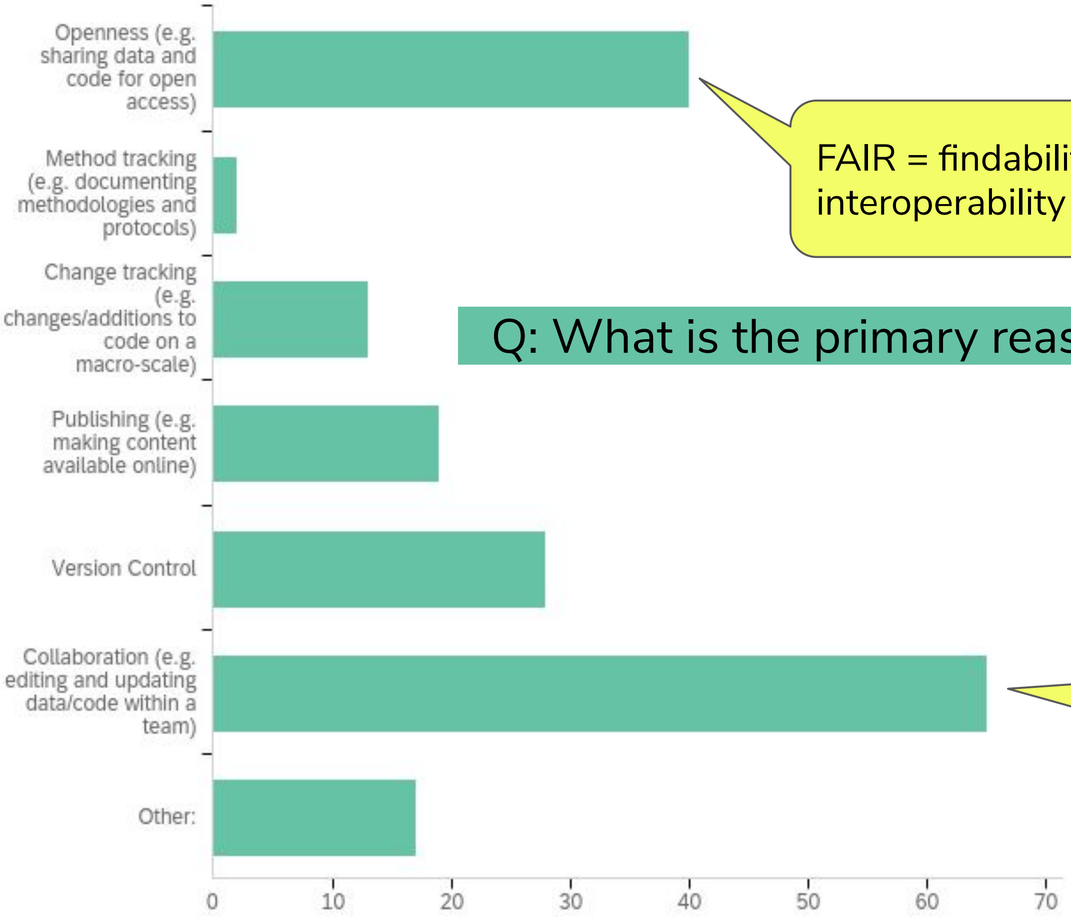


#IASGE

Survey - Preliminary Findings

Please participate & share widely!
bit.ly/3aO0ykQ




Openness (e.g. sharing data and code for open access)
Method tracking (e.g. documenting methodologies and protocols)
Change tracking (e.g. changes/additions to code on a macro-scale)
Publishing (e.g. making content available online)
Version Control
Collaboration (e.g. editing and updating data/code within a team)
Other:
0
10
20
30
40
50
60
70

FAIR = findability, accessibility, interoperability and reusability

Q: What is the primary reason you use git hosting platforms?

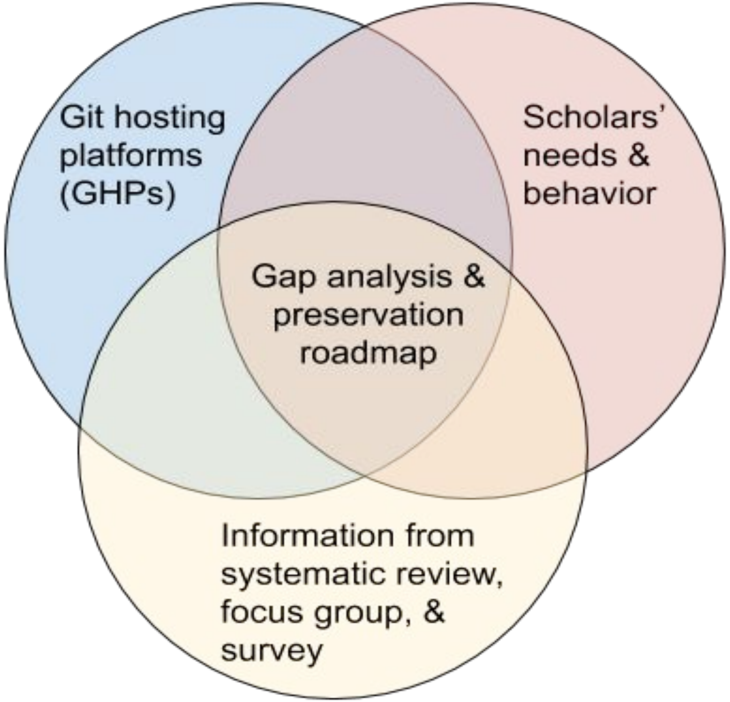
Community, collaboration, & peer production

#IASGE

Part IV: User Experience Interviews (Qual)

Semi-structured interviews with 50 scholars to understand their behaviours

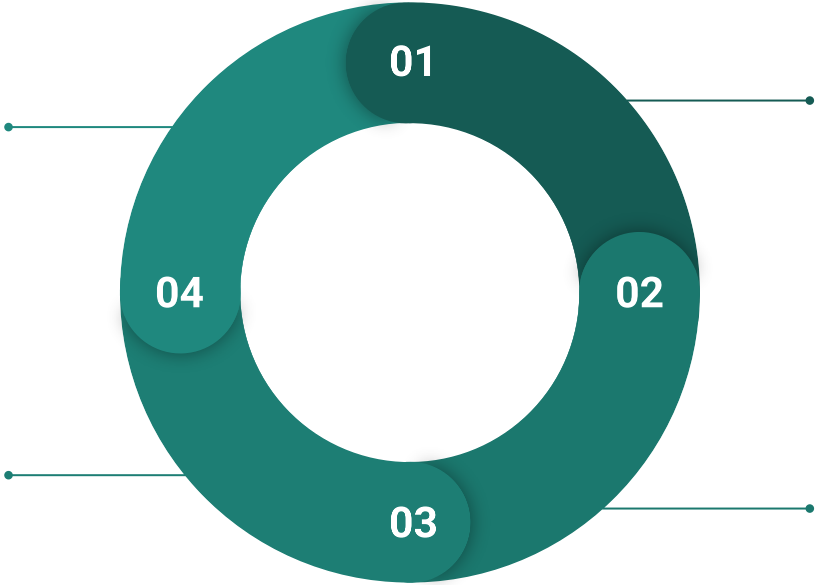
- Why did folks stop using Git & GHPs?
- If/how are they versioning their work without these toolkits?
- What features & workflows are in practice that we have not heard of before?


Git hosting platforms (GHPs)
Scholars' needs & behavior
Gap analysis & preservation roadmap
Information from systematic review, focus group, & survey

Current Archival Approaches

#IASGE

IASGE Environmental Scan



Software Preservation

Best practices, software curation and description

Programmatic Captures

large-scale archiving of GitHub API data; large-scale archiving of source code from Git hosting platforms; select archiving of repos

Web Archiving

State-of-the-art web archiving tools and technologies; Could they be used for software capture?

Self-Archiving

Motivations to self-archive; appeal of general repositories vs. institutional repositories

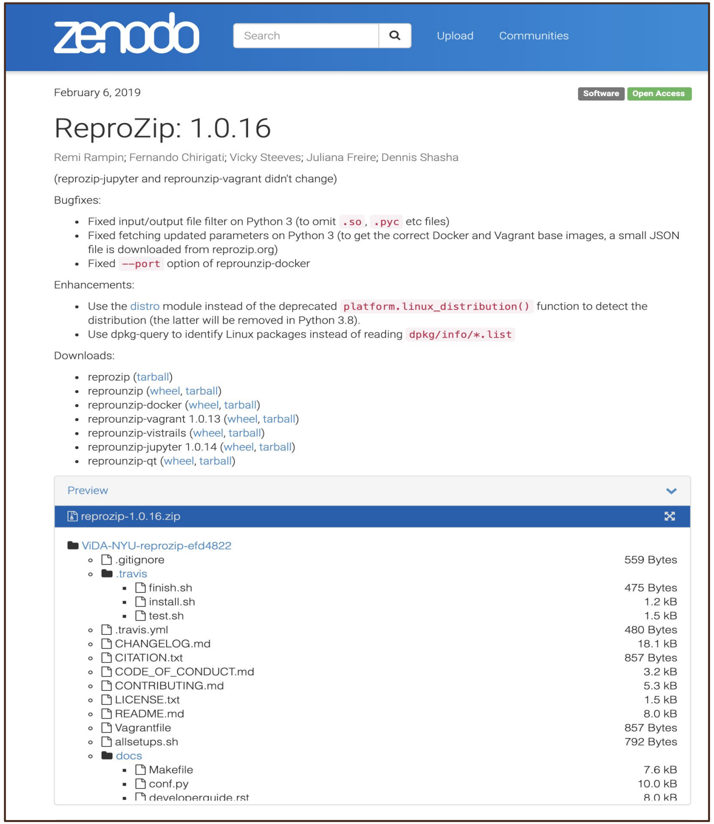
Web Archiving

- The use of web crawlers/software to capture web-based content
- Understand who, if anyone, is currently using this technology for code repositories

- Projects of Interest:
  - Archive-It
  - Webrecorder
  - Memento Tracer

Self-Archiving

1. Identify motivations for self-archiving in general repositories
2. Identify gaps regarding IRs and hosting/describing software
3. Understand platform integration (GitHub <> Zenodo)
4. Should we keep rehabbing IRs or move to more flexible models?



February 6, 2019

Software Open Access

ReproZip: 1.0.16

Remi Rampin; Fernando Chirigati; Vicky Steeves; Juliana Freire; Dennis Shasha

(reprozip-jupyter and reprounzip-vagrant didn't change)

Bugfixes:

- Fixed input/output file filter on Python 3 (to omit `.so`, `.pyc` etc files)
- Fixed fetching updated parameters on Python 3 (to get the correct Docker and Vagrant base images, a small JSON file is downloaded from reprozip.org)
- Fixed `--port` option of reprounzip-docker

Enhancements:

- Use the distro module instead of the deprecated `platform.linux_distribution()` function to detect the distribution (the latter will be removed in Python 3.8).
- Use dpkg-query to identify Linux packages instead of reading `dpkg/info/*.list`

Downloads:

- reprozip (tarball)
- reprounzip (wheel, tarball)
- reprounzip-docker (wheel, tarball)
- reprounzip-vagrant 1.0.13 (wheel, tarball)
- reprounzip-vistrails (wheel, tarball)
- reprounzip-jupyter 1.0.14 (wheel, tarball)
- reprounzip-qt (wheel, tarball)

Preview

reprozip-1.0.16.zip

- ViDA-NYU-reprozip-efd4822
  - .gitignore 559 Bytes
  - .travis
    - finish.sh 475 Bytes
    - install.sh 1.2 kB
    - test.sh 1.5 kB
  - .travis.yml 480 Bytes
  - CHANGELOG.md 18.1 kB
  - CITATION.txt 857 Bytes
  - CODE_OF_CONDUCT.md 3.2 kB
  - CONTRIBUTING.md 5.3 kB
  - LICENSE.txt 1.5 kB
  - README.md 8.0 kB
  - Vagrantfile 857 Bytes
  - allsetups.sh 792 Bytes
  - docs
    - Makefile 7.6 kB
    - conf.py 10.0 kB
    - developerguide.rst 8.0 kB

Software in Zenodo with Integrations with GitHub and indexed by OpenAIRE

#IASGE

Programmatic Captures

1. Software to capture software
2. The use of indexing, cloning, and APIs to capture software and contextual information
3. Scholarly ephemera in one place, software in another
4. Projects of interest
   a. Software Heritage
   b. GHTorrent and GH Archive
   c. SARA

SARA

SOFTWARE ARCHIVING OF RESEARCH ARTEFACTS

git GH Archive

Software Heritage

PRESERVING TECHNICAL KNOWLEDGE

The GHTorrent project

Software Preservation

1. Understanding current communities of practice
2. Software metadata & citation
3. Software curation
4. Projects of interest
   a. Software Preservation Network
   b. Software Emulation (e.g. EaaSI)
   c. Software Sustainability Institute
   d. US Research Software Sustainability Institute

Software
Sustainability
Institute

BETTER
SOFTWARE
BETTER
RESEARCH

EaaSI

Software
Preservation
Network

Towards an Archival Spec

Culmination of this research will be an archival spec that can be used by institutions.

Details will include:

1. The capture of both source code and its ephemera (commit messages, merge requests, issues, wikis)
2. Description and curation
3. Sample preservation workflows

WA Tool Testing Phase - Preliminary Findings

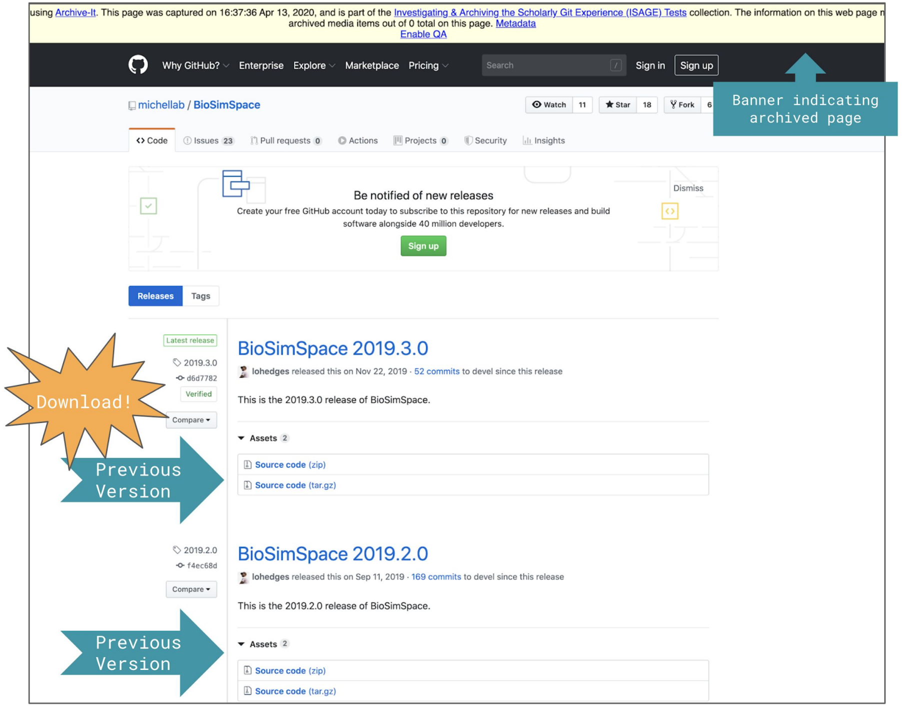
Archive-It

a. Test crawl four git repositories using the standard crawler and Brozzler

b. Found issues with capturing GitLab (rendering issues) and Bitbucket (only got a white page!)

c. Promising results with standard crawler on GitHub

d. Limitations: patch crawling issues, time needed, and “uncharted territory” using this method



using Archive-It. This page was captured on 16:37:36 Apr 13, 2020, and is part of the Investigating & Archiving the Scholarly Git Experience (ISAGE) Tests collection. The information on this web page m
archived media items out of 0 total on this page. Metadata
Enable QA

michellab / BioSimSpace

Be notified of new releases

Create your free GitHub account today to subscribe to this repository for new releases and build software alongside 40 million developers.

BioSimSpace 2019.3.0

lohedges released this on Nov 22, 2019 · 52 commits to devel since this release

This is the 2019.3.0 release of BioSimSpace.

BioSimSpace 2019.2.0

lohedges released this on Sep 11, 2019 · 169 commits to devel since this release

This is the 2019.2.0 release of BioSimSpace.

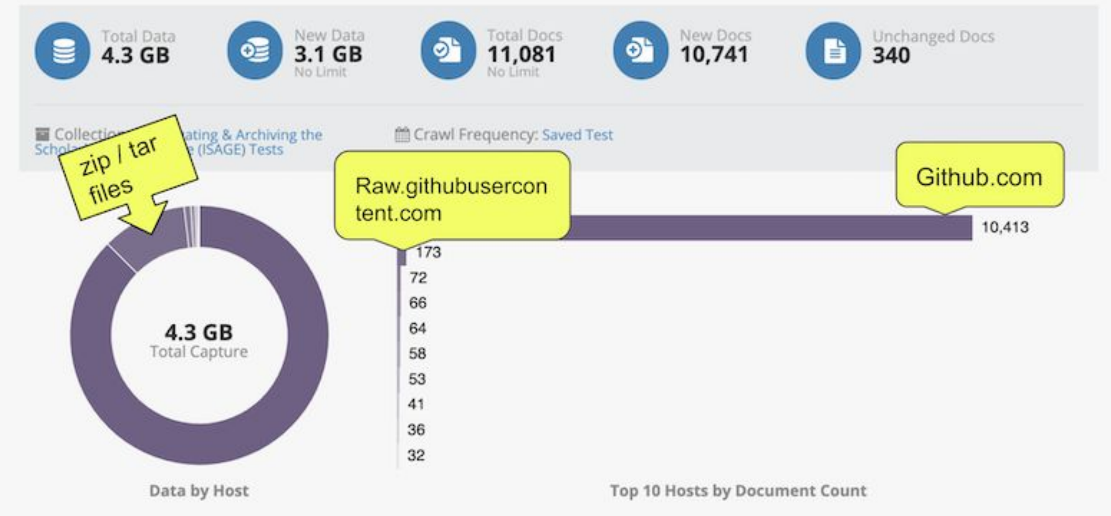
Read more: Lab Notes detailing Archive-it testing:
investigating-archiving-git.gitlab.io/updates/lab-notes-archive-it

#IASGE

WA Tool Testing Phase - Preliminary Findings

What was captured:

- Zip of source code with past versions
- 23 open issues (at time of capture) and 111 closed issues, and their labels.
- PRs and messages were also captured, except for indiv. Commit messages
- 4 pages of commits were saved, but older commit messages were missing



Total Data 4.3 GB

New Data 3.1 GB No Limit

Total Docs 11,081 No Limit

New Docs 10,741

Unchanged Docs 340

Collection: ...ating & Archiving the Schol... ...e (ISAGE) Tests

Crawl Frequency: Saved Test

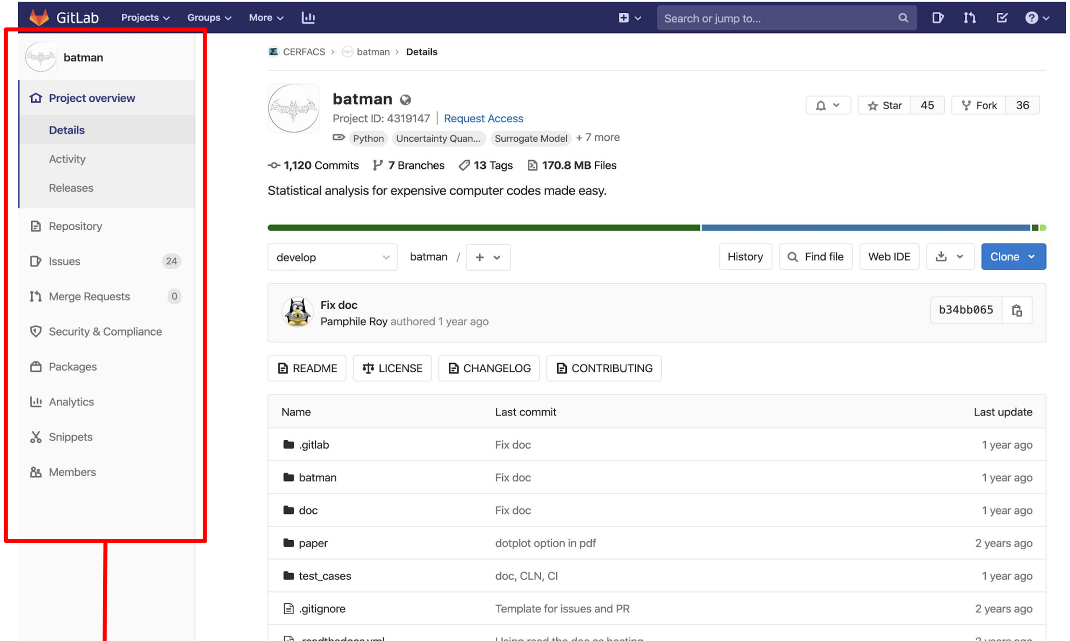
Screenshot of Archive-It crawl report using Standard crawler for single GitHub repository link from: investigating-archiving-git.gitlab.io/updates/lab-notes-archive-it/

#IASGE

What’s next? Further testing!

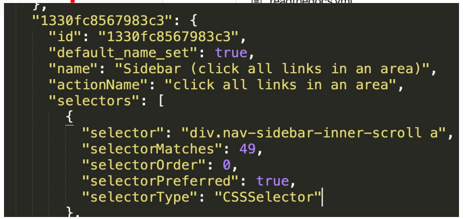
- **Memento Tracer**
  - JSON traces serve as “instructions” for the crawler, which can be shared and reused
  - JSON file can be applied to any repositories on that domain (i.e. any repository on GitHub).
  - http://tracer.mementoweb.org/

- **Webrecorder**
  - Uses user interactions (clicking, scrolling) to create high quality captures
  - Limited scalability
  - Potential to create an Autopilot behaviors for GitHub? feature request ;-)
  - https://webrecorder.io/



Statistical analysis for expensive computer codes made easy.

Name	Last commit	Last update
.gitlab	Fix doc	1 year ago
batman	Fix doc	1 year ago
doc	Fix doc	1 year ago
paper	dotplot option in pdf	2 years ago
test_cases	doc, CLN, CI	1 year ago
.gitignore	Template for issues and PR	2 years ago
.readthedocs.yml	Using read the doc as hosting	2 years ago



```
},
"1330fc8567983c3": {
  "id": "1330fc8567983c3",
  "default_name_set": true,
  "name": "Sidebar (click all links in an area)",
  "actionName": "click all links in an area",
  "selectors": [
    {
      "selector": "div.nav-sidebar-inner-scroll a",
      "selectorMatches": 49,
      "selectorOrder": 0,
      "selectorPreferred": true,
      "selectorType": "CSSSelector"
    },
```

JSON trace created with Memento Tracer Chrome extension

#IASGE

Summary, calls to action, & all our contact info!

- Code/software and the contextual ephemera are worth saving
- Scholarship in Git format & in Git hosting platforms is at risk because there is no preservation plan
- Understanding behaviour patterns of academics using GHPs will help in T&L and archiving work
- Be on the lookout for research-centered blog posts
- Send us feedback on posts and/or resources you think we should know about!

Project website:
https://investigating-archiving-git.gitlab.io

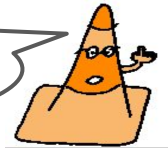
GitLab repo:
gitlab.com/investigating-archiving-git

Emails:
vicky.steeves@nyu.edu
genevieve.milliken@nyu.edu
sarahtnguyen@nyu.edu

Twitters:
@VickySteeves
@gen_milliken
@snewyuen

Survey - Preliminary Findings

Please participate & share widely!
bit.ly/3aO0ykQ



Summary

Q33 - Do you use git hosting platforms as a storage place to backup your code?

- % said Yes

Q37 - How is your research or scholarship funded? Check all that apply.

- Mostly public

Q42 - How often do you use git to collaborate on authoring code?

- Mostly daily

Q43 - When a new collaborator joins your team, is there an onboarding process or protocol specifically for introducing them to your coding practices and use of version control?

- Mostly No

Q47 - Do you copy your repositories to external long-term storage services or platforms (e.g. Zenodo, OSF, institutional repository)?

- Mostly No

#IASGE