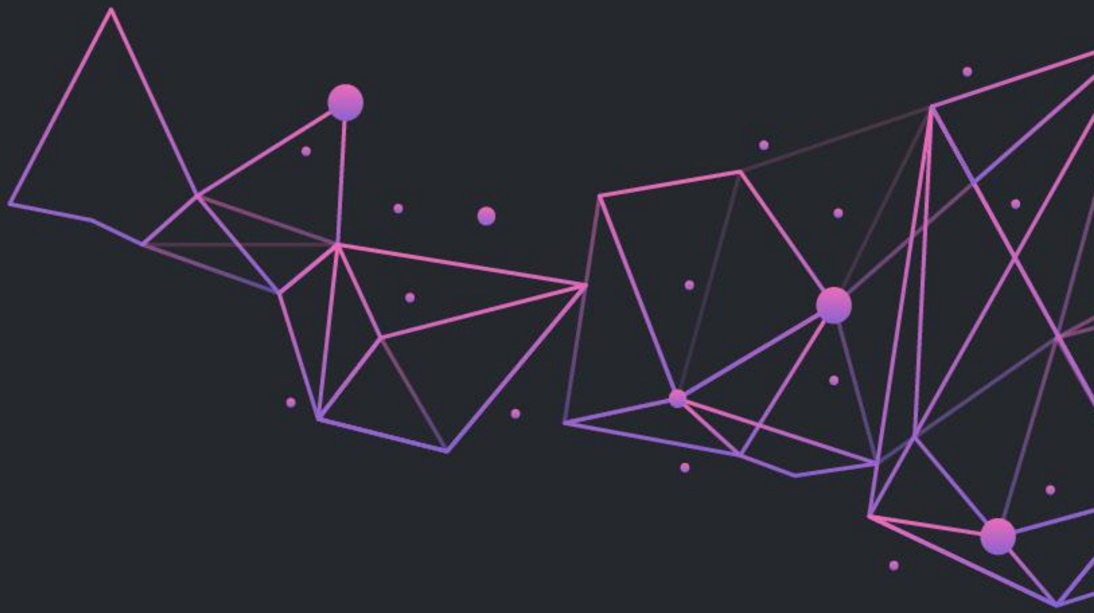
For Public Release

Report on standardized GitHub metrics for international development, public policy, and economics



Based on the report to GitHub by Tattle Civic Tech

Last updated: 1 September 2023

Tattle Civic Tech ("Tattle") was selected to lead the research project, and delivered the research outputs that formed the basis of this report. The report was written by Tarunima Prabhakar, Co-Founder and Research Lead of Tattle, and Mala Kumar, former Director of Tech for Social Good at GitHub. Upasana Bhattacharjee, Researcher at Tattle, and Swastika Mohapatra, Data Science Intern at Tattle, were contributing authors. Contributing editors were Yash Budhwar, Operations and Communications Lead at Tattle, and GitHub staff: Cynthia Lo, Skills-Based Volunteering Program Manager; Peter Cihon, Senior Policy Manager; Abby Rieflin, Senior Corporate Communications Manager; Caffrey Lynch, Senior Corporate Communications Specialist. Special thanks go to Kevin Xu, Senior Software Engineer at GitHub, as well as the many international development, public policy, economics, and academic experts who were interviewed as part of the research.

Executive summary

There is increasing demand for data that contextualizes the impact of software collaborations across geographies. GitHub is the largest global platform for code collaboration, and thus data aggregated over its public activity is uniquely insightful for researchers and practitioners working in related fields, such as international development, public policy, and economics. This report maps the data needs of researchers and practitioners in these fields to the data sharing possibilities at GitHub. In order to better understand the needs of the community, GitHub engaged Tattle to discover how private sector organizations can support or augment public sector data and to share these findings. The authors realized this research report can also provide guidance to corporations looking to share data.

While there are potential risks of GitHub sharing data, some of which are detailed in this paper, the authors find a strong need for GitHub to responsibly release aggregated metrics,[1] which can serve as an important resource for understanding the impact of software and data-centered projects on global development. In absence of data released by GitHub, researchers will need to rely on third-party services, some of which may not be up-to-date or complete.

The authors observe significant diversity in the questions, topics, methods, and resources of analysis among potential users of these metrics. Specifically for the international development community, aggregate but stable metrics at a national level are sufficient. Academic researchers generally requested more granular levels. The authors emphasize the importance of GitHub establishing consistent, transparent, and consultative processes through which the metrics are conceptualized and released.

This report contains the background, major findings, and recommendations for the research project that GitHub has published, the GitHub Innovation Graph. The GitHub Innovation Graph was informed by this report and provides downloadable CSV files that are generally publicly available and updated regularly. The GitHub Innovation Graph specifications help establish consistent definitions and units of measurement to establish benchmarks to measure progress towards the Sustainable Development Goals (SDGs) within a country and between countries. As ‘data deserts’ negatively impact communities that data sets omit particularly in low-and-middle-income countries, the GitHub Innovation Graph aims to provide insights that can inform public policy and international development priorities to help opportunities for action.

The GitHub Innovation Graph is the most comprehensive attempt to represent software developers and software-adjacent technologists toward the advancement of the Sustainable Development Goals (SDGs).

[1] In this paper, “data” refers to individual GitHub user activity or attributes, such as user location. “Metrics” refers to cleaned, aggregated GitHub data. See the key terms section for more information.

Key terms

Big data

High volume of unstructured, semi-structured, or structured data.

DPG/Digital Public Goods

"Open source software, open data, open AI models, open standards and open content that adhere to privacy and other applicable laws and best practices and do no harm and are of high relevance for attainment of the United Nations 2030 Sustainable Development Goals (SDGs)."[2]

ICT/information and communication technologies

A broad term to refer to infrastructure that enables digitalization, including landline telephones, mobile devices, software applications, and media devices.

Indicators

Indicators can be used to drive a decision through comparison with previous values. An indicator may be a composite of several metrics.

GitHub activity/activity on GitHub

Activity refers to the 50+ event types that are counted and/or trigger an action on GitHub that can be tied to a user or organization. Activities include but are not limited to pushing code, adding a comment to an issue, and editing a markdown file. Refer to GitHub documentation for a comprehensive list of event types.[3]

LMICs/ Low- and middle-income countries

According to the World Bank, LMICs are countries in the lower three of four brackets, as defined by gross national income per capita.[4]

Metrics

Metrics are aggregated data that on their own, do not imply strength, value, or conditions that drive action. Examples of metrics are the population of a country or the number of commits in a GitHub repository.

Open source software

[2] https://digitalpublicgoods.net/standard/

[3] https://docs.github.com/en/developers/webhooks-and-events/events/github-event-types

[4] https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Section: Key Terms

The open source ecosystem extends to many types of products and data. The working definition of open source software (OSS) used in this report is software distributed and uploaded to an accessible repository with source code that may be read or modified by users, and that is under an Open Source Initiative OSS license.

Open content and open data

OpenDefinition.org states, “Open data and content can be freely used, modified, and shared by anyone for any purpose.” As is the case with OSS, open data must be publicly available, in a machine-readable format, and easily accessible online.[5]

Platform

In this report, we are specifically referring to digital platforms or software-based infrastructure that facilitates interactions between users.

Public code

Code, licensed or unlicensed, that is on a public repository.

SDGs/Sustainable Development Goals

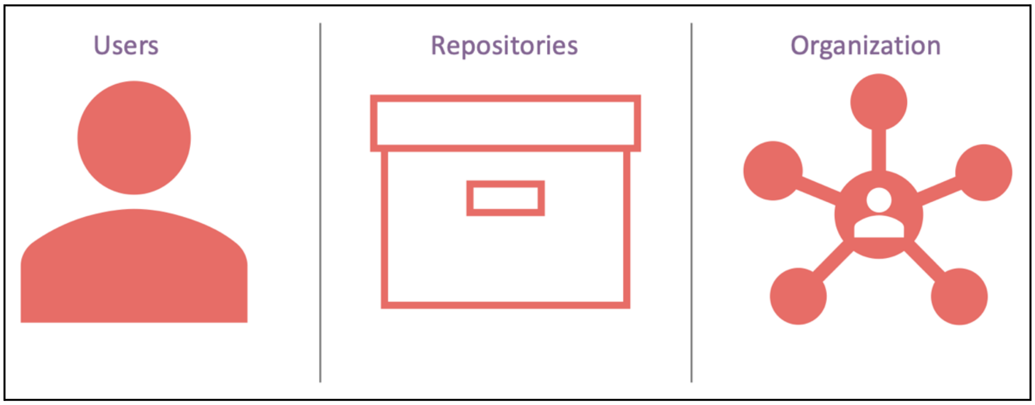
Seventeen goals set by the United Nations along thematic areas of human progress, such as education, food security, and public health. Each goal has multiple targets and indicators.

[5] http://opendatatoolkit.worldbank.org/en/essentials.html

Activity on GitHub

GitHub is a platform that connects users to coding projects.[6] Users are individuals and may belong to organizations and collaborate on project repositories. A user or an organization may create or own one or more coding projects, which may have one or more repositories. All activity on GitHub describes some relationship between users, repositories, and organizations.

Figure 1: The three main entities on GitHub



There are different ways in which these three entities can be connected. For example, a user may own a coding project or create a repository for that coding project. They might fork (copy) a repository, commit (add) to a repository, create pull requests to make a suggestion, star a repository to add it to their favorites, or create an issue to describe a problem. Users might follow other users or organizations. All these activities reflect a varying level of engagement with a specific project repository.

User, repository, and organization activity, and the interaction therein can inform a range of research questions.

[6] Used in the general sense of the word and not in reference to the projects feature on GitHub.

Introduction

The ongoing digitalization of the economy and society relies heavily on all forms of software collaboration activity, be it open source, public, or private. GitHub is the world's largest platform for code collaboration, and its usage depends on numerous economic and societal factors, such as literacy, wealth, and internet accessibility.

Over the last two decades, data about digital platforms usage has been increasingly used as an important public policy and programmatic tool to achieve the Sustainable Development Goals (SDGs). Since platform data is driven by a different set of institutional origins, practices, and biases, it can be used to cross-validate existing and older methods of collecting population wide parameters (Taylor & Schroeder, 2015). In absence of on-ground data collection exercises by governments or multilaterals, data from digital platforms can also fill longstanding data gaps.

This report maps the data needs of researchers and practitioners in these fields to the data sharing possibilities at GitHub. In this paper, we focus specifically on how GitHub can release metrics to help international development, public policy, and economics researchers and practitioners make better-informed decisions. Releasing the GitHub metrics described in this paper is the most comprehensive attempt to represent how software developers and software-adjacent technologists contribute to the advancement of the SDGs. We define metrics as *data aggregated, across the platform, and over several projects*.

Based on interviews with professionals in international development, public policy, and economics, the authors propose a list of conceptual metrics that GitHub can release to inform the research questions identified around:

- The location of GitHub users and projects
- Public code collaboration participants
- The motivations and roles of GitHub users
- The health of and the governance of projects related to the SDGs
- GitHub project and user journeys over time
- Factors that result in the growth of public code collaborations, especially for social good

The next sections include project background and a discussion on the risk of "big data." Readers can go to the major findings and recommendations sections for the research project outcomes.

Project background

In October 2021, the GitHub Social Impact and Policy teams released a request for proposals (RFP) to inform how to responsibly release aggregated data on its platform activity to aid international development, public policy, and economics research and programs. Tattle Civic Tech was selected as the vendor and submitted its final reports to GitHub in September 2022.

Research methodology

Tattle used three main research inputs to inform their reports:

- A detailed literature review on previous papers that used data about GitHub activity
- A Google form to solicit volunteers interested in participating in the research project
- Semi-structured, detailed interviews

Through the Google form and literature review, Tattle identified potential volunteer interviewees who had extensive experience in working with GitHub activity data and/or working with other digital platform data in international development. Tattle conducted one-on-one interviews through Zoom with 18 researchers and practitioners working in international development, public policy, economics, technology, and open source ecosystem development. Geographically, the interviewees came from North America, Europe, Africa, and Asia. Professionally, the interviewees spanned roles such as program managers, data scientists, academic researchers, engineering managers, and strategists. Tattle also interviewed employees from different teams at GitHub to understand the internal considerations on releasing metrics.

Depending on the interviewee's professional background, the interview covered one or more of the following:

- Experience and challenges in working with data, from digital platforms or otherwise
- Prior experience with GitHub and data about GitHub to understand motivations, points of friction, and recommendations for others
- Questions of interest to the concerned stakeholder that could be helped with GitHub metrics
- Data use practices to understand how GitHub should share the metrics

Tattle then consolidated findings from the interviews and conducted two focus groups to get more targeted feedback. The first focus group was with researchers familiar with GitHub as a platform. Many of them had used GitHub activity data in their work. The second focus group was with international development practitioners. These focus group discussions also included eight individuals working in international development who had not been previously interviewed. This gave the research team an opportunity to evaluate the relative importance of different metrics for different groups.

The feedback from the focus group discussions was consolidated into two reports that Tattle submitted to GitHub, which GitHub then used to write this paper.

Scope

Most of the research studies analyzed in the literature review were interested in the collaboration dynamics of GitHub projects and were therefore using project-level data. The scope of this research project, however, is metrics from activity across the entire GitHub platform, not individual projects. This report includes a section for the requests received for project-level metrics that GitHub may consider in the future.

This report summarizes the concerns and needs expressed in the research, as well as the reputational and business considerations for GitHub in releasing metrics to support those working in international development, public policy, and economics. Assumptions and concerns came from three primary groups:

- Those focused on the opportunities of open data from digital platforms
- Those outside of GitHub focused on minimizing harm to individuals from identification in datasets
- Those at GitHub focused on minimizing harm to the organization and GitHub users

This report balances the interests and concerns of the groups by assuming that the maximal opportunity comes with responsible data sharing. This report also attempts to document and communicate the perspectives of the groups to each other.

Risk analysis of GitHub metrics conceptualization

Given that it is difficult to foresee the (mis)uses of releasing information to the public, a risk analysis that explores the potential benefits against the potential harms should be undertaken as the metrics, including the level of aggregation, are conceptualized. The effects of cross-linking external datasets for the possible identification of users, plus the possible harm mitigation strategies, should be considered ahead of time.

Possible harms can have several dimensions. Based on MERL Tech's work and existing research on online harms (MERL Tech, n.d; Agrafiotis et al, 2018), the following dimensions of harms should be considered as datasets are made public:

- Physical: physical and bodily injury
- Legal: profiling, repression, impact on legal and fundamental rights
- Economic: financial losses or negative economic consequences
- Psychological or emotional: emotional and psychological distress, impact on mental well-being
- Social/Societal: broader harms on a societal level, perception, and organization
- Reputational: damaged public perception and goodwill
- Other: additional and/or context-specific harms

Below are the various stakeholders who will be affected by data sharing by GitHub, and possible risks:

Table 1: Possible risks by stakeholders in releasing GitHub metrics

Stakeholder	**Possible risks**
National and sub-national governments	Metrics might be used to evaluate national performance, which could affect the reputation of government bodies.
Technology companies	Metrics could affect the perception of specific open source technologies, or the actions of companies that work on or with public code.
Open source communities – international, regional, local/grassroot	Metrics could affect funding decisions or bring unintended bias or scrutiny to open technology communities working on sensitive issues.
GitHub users	If identified, GitHub users could be targeted based on their gender; sexuality; occupation; education; location; race, caste, ethnicity; spoken language; age; dis/abilities; religion; socioeconomic status, or an intersection of these demographics.
International development sector	Metrics specific to sustainable development projects could affect evaluations of the work done by international development organizations.
Academics and researchers	Academics and researchers are primarily consumers of the metrics. We do not foresee possible direct harm to this stakeholder.

Section: Discussion points around "Big Data"

GitHub (the company)	The metrics could increase scrutiny on GitHub and affect its business interests.

The EU data protection authorities list nine criteria[7] that indicate high risk in processing data, one of which is matching or combining datasets. Pozen (2005) pointed out that disparate items of information, though individually limited or of no utility to their possessor, can take on added significance when combined with other items of information, creating a 'mosaic effect'. In a famous case, researchers were able to compare an anonymized dataset of customer movie rankings from Netflix with public information in IMDB. They identified Netflix consumers and sensitive details, such as their political preference (Narayanan & Shamtikov, 2008).

One hypothetical scenario with GitHub metrics involves "the number of projects by development pathway, in a region". As an isolated dataset, this would help researchers and international development agencies to understand interest and investments. If this were mapped onto external location data and aggregated data on popular projects by geography, however, GitHub data might be used to identify granular project hotbeds. Were the project not favorable to governments or other local populations, activity might be curtailed through internet shutdowns or intimidation.

On the other hand, data from digital platforms should not be used in isolation from local contexts where GitHub users are based. While data from platforms can be used as a proxy for missing statistical data, especially in low-income countries, the absence of survey data can indicate a lack of resources.[8] These constraints may affect the verifiability of big data, because platform data may not be cross-checked through on-the-ground surveys.

Relying on platform data to inform policy decisions and initiatives can also shift the decision-making power over data from countries and governments to corporations (Taylor & Schroeder, 2015). This increases corporate responsibility to consider the interests of all their users. Arora (2016) expressed concern that corporations have a legacy of treating populations in the low- and middle-income countries (LMICs) as new customers being sold dreams of empowerment without having any power to negotiate or control their data. Researchers and practitioners from the LMICs might be disproportionately excluded or given insufficient recognition of their work or the work of their institutional affiliations. Leaning towards more open datasets can allow for a thriving and more diverse research community. This also prevents redundancy in efforts of data collection by different groups, which is especially important in resource-constrained areas.

Researchers have also warned against data-driven solutionism, where oversimplification and misrepresentation creeps into the identification, definition, and construction of societal problems (Morozov, 2013). The onus of responsible and appropriate use of data is shared by the entity releasing the data (in this case GitHub), as well as the researchers using the data. From this perspective, the consultative process undertaken by GitHub is important. In addition to transparency around how the metrics are conceived, it is also important to highlight how and when the metrics from GitHub should not

[7] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias/when-do-we-need-to-do-a-dpia/#when2

[8] Active data is explicitly requested from the user. Surveys are a common method of producing active data.

be used. This is the approach taken in conceptualizing the metrics (see the Recommendations section for more details).

Major findings

Personas

The authors found significant heterogeneity in how researchers and practitioners in international development, public policy, and economics engage with research questions and with different methods for accessing and processing data. The researchers and practitioners also had a wide range of resources and capacities, in terms of connectivity, computing power, and labor. These must factor into which metrics are released and how.

Interviewees were classified into the following three broad personas:

Persona 1: Academics studying collaboration, collaborative platforms, and innovation

- Have a clear idea of the metrics that would be useful to their research and their implications
- Metric requirements are precise and often granular, like:
    - Number of commits from a region
    - User contributions to projects
    - Location distribution of users/projects
- Have access to strong computing power and skilled talent to work with data

Persona 2: International development professionals

- Metrics tend to be instruments to inform work in development, infrastructure assessment, project development, innovation, and more on sustainable development and civic issues
- Less likely to know exactly how GitHub metrics can support their work.
- The metric requirements are more abstract and aggregate, with a focus on, access, geography, and government data, including:
    - Internet adoption per country
    - Tech literacy by country/region/city
- Less likely to have access to computing power and skilled talent to work with data

Persona 3: Community leaders/members working on sustainable development and civic issues

- Metrics aid in tracking the success of collaboration/contributions, interactions, and inclusivity of their open knowledge communities
- Demographic data, although difficult to collect, is valuable, including:
    - Diversity in projects
    - Developer interactions
    - Skill development
- Often resource-starved and need data to tell stories about their impact

Each persona was found among people working in international development, public policy, and economics, though some personas were more commonly found in one of the three.

Academics were more commonly involved in conceptualizing indices around economic growth. International development practitioners were interested in using such indices. Community leaders of collaborative and/or open tech projects (Persona 3) could be working towards public policy goals or in international development.

Domains of interest

Tattle attempted to uncover what aspects of users, projects, organizations, and their interactions were of interest to the three personas. This section details the questions that surfaced in the research, without consideration as to whether they can be answered by data from GitHub. The questions are classified under the 5W (and 1H) framework and in the rough order of importance, as determined by how frequently an interviewee said it was used or expressed interest in this information.

Where

Where public code collaboration takes place was one of the most important topics. Some of the broad geography-related questions identified were:

Table 2: Questions and discussion points around the “where” of releasing GitHub metrics

Questions	Discussion points
Where are users located?	GitHub activity reflects a more sophisticated and technically proficient engagement than social media activity. Thus, the location of GitHub users can serve as a proxy for tech proficiency, productive skills, and internet availability and shutdowns in a region.
Where do projects originate?	A coding project could include one or more repositories tied to an individual or an organization. As described in a later section, trying to determine a specific location is not straightforward. Regardless, the location of a project can convey the needs and technical ethos of a place.
What is the geographical relationship between the users and the projects to which they contribute?	As a global platform for code collaboration, GitHub enables internet users to discover and work on projects from around the world. What does the user location and the project location convey about information and capital flows? For example, are projects started in high-income countries, with contributors from LMICs joining later? Or is it the other way around? Where are the active contributor hubs?

Who

Researchers are interested in knowing more about who is on GitHub. The ‘who’ could concern user demographics (gender, spoken language), professional affiliation (student or employer), or their role on a project. They are also interested in whose accounts could be affected by government action or targeted attacks.

Section: Major findings

'Affiliation' in the context of public code collaboration can mean many things. People may contribute to public code projects through the organization that provides them full-time employment. They might do it while volunteering at an organization outside of their full-time job. They might do both. People might also have multiple professional affiliations on the platform. Researchers are interested in understanding the social and economic backgrounds of users that enable a person to contribute to public code.[9]

What/why

Researchers are interested in projects classified by SDG thematic areas, such as health, education, and food security. Humanitarian aid organizations are interested in identifying robust projects and their funding and organizational structures to support the SDGs. They might also be interested in comparing the 'health'[10] of different projects. Finally, to differentiate between open source within public code collaborations, researchers are also interested in the license terms under which different projects on GitHub are released.

When

Linking GitHub activity to time can help show the evolution of projects, as well as GitHub user journeys. For example, how many projects stay active over a five-year period? How, if at all, does open source participation contribute to skilling? How many 'power users' remain high contributors on a project? Or how does GitHub activity change when users graduate from universities into professional settings?

How

A complex area of research is how public code/open source collaboration takes place. For example, how do non-coding and coding roles interact with each other? What leads some projects to become popular and stable, and others to stall? How does financing shape social good projects?

Questions Summary

The table below summarizes the research questions and connects them to the user personas that are most interested in these questions. Not all questions can be answered through metrics aggregated over all public activity on GitHub. Answering some of these requires metrics at the project level. In the Recommendations section, we list the aggregate metrics that GitHub can release to inform these questions, as well as project-level data that GitHub can consider as a part of subsequent projects.

Classification	**Questions**	**Interested personas**

[9] In April 2020, GitHub published a report that covers the economic and social factors around contributing to open source for social good projects. Download the report for free here: https://socialimpact.github.com/assets/img/research/GitHub_tCF_OSSInSocialSector_FINAL_updated.pdf

[10] See: Xia, T., Fu, W., Shu, R., Agrawal, R., & Menzies, T. (2020). Predicting Health Indicators for Open Source Projects (using Hyperparameter Optimization). doi:10.48550/ARXIV.2006.07240

Where	● Where are users located? ● Where do projects originate? ● What is the geographical relationship between the users and the projects they contribute to?	Academics, professionals in international development; open tech community leaders and members
Who	● Who is on GitHub? ● What are their demographics (gender, spoken language), and professional affiliation? ● What role do they play in projects?	Academics and open tech community leaders/members
What/Why	● What thematic area within SDGs are projects focused on? ● Under what license terms are projects released? ● Why do users contribute to certain projects, and what kinds of projects receive the most contributions?	International development professionals and open tech community leaders/members
When	● How do projects evolve over time? ● How do user journeys evolve over time?	Academics
How	● What factors contribute to the growth of public code collaborations?	Academics, professionals in international development, and open tech community members

Data release as perception management

There was significant diversity in how interviewees perceived GitHub. For some, GitHub is the primary platform for open source collaboration. They had positive perceptions of the platform. Some emphasized GitHub's nature as a corporate entity that relies on enterprise customers for revenue. Others were wary of GitHub's connection to Microsoft. Some interviewees expressed doubts about GitHub ever releasing datasets for research. A few were suspicious about GitHub's intentions in participating in open access initiatives (this project included). In one case, an open tech community leader declined an interview since they didn't believe in GitHub's commitment to open source processes.

The general perception is that GitHub has not released metrics for research, even though GitHub has released some metrics in its annual State of the Octoverse reports.[11] As per one researcher tracking software development and open source processes:

> "There is no actual GitHub dataset. GitHub does not officially sponsor a dataset. The datasets that are available are all built and maintained by third parties. The way I understand it, if right now you

[11] https://octoverse.github.com/

> want to use GitHub data, you either work within the API based on the rate limits that are available and you sample across different kinds of public instances."

The researcher went on to say that third-party aggregated alternatives, such as GH Archive or GHTorrent, are underfunded and that some researchers are building their own mirrors, which is computationally inefficient. They also described how the GitHub REST API and Events API are for developers to "create integrations, retrieve data, and automate workflows," not for research purposes, which create sampling bias issues.

One interviewee tempered their enthusiasm around more platform data. They warned that while many researchers and practitioners believed that more data would be useful to their work, how the data would be used wasn't always clear when making the request. Further, they cautioned that GitHub users generally do not expect data about their platform activity will be used by researchers. Lastly, regulators might not understand the difference between GitHub and a social media or e-commerce platform. Data released by GitHub could bring it more media and ultimately regulatory attention. In recognition of these concerns, the authors included a Risk Assessment (see above) in this paper.

The opportunity of releasing data for social impact–driven research can also be seen as managing GitHub's perception amongst diverse external stakeholders. While there are consequences for sharing data that need to be thought through, there are also consequences for *not* sharing data. Releasing metrics responsibly would reinforce GitHub's commitment to platform transparency, as well as social impact and open access initiatives.

Metrics releases

Data about GitHub activity is currently being used in indices and reports in many disciplines. In absence of one authoritative GitHub data source, researchers rely on third-party sites, such as GH Archive and GHTorrent, which provide different data points and are updated at different frequencies. The data from these sources might not reflect recent usage and activity. Interviewees mentioned that the consumers of these indices are usually decision-makers who do not have the time or the willingness to verify the accuracy or data aggregation methods of the data. In absence of alternatives, the data from external sources may incorrectly be perceived to be authoritative. This is both a reputational and business risk to GitHub.

By owning the data about GitHub, the company can provide the necessary context on how it was generated and known data voids. It will also enable a standardization of data used externally. Finally, it provides one authoritative source to consult when trying to verify and validate GitHub data.

When it comes to international development, simple is good

For those in international development (persona 2), broad aggregate metrics at the country (national) level released annually or quarterly are useful. Stability of metrics over multiple years is important for this persona since they are interested in evaluating improvements over baseline metrics.

Watch out for the bots!

One of the interviewees highlighted that bots on GitHub that automate Actions (workflows), can present a challenge in computing metrics. Bots should be filtered out of the metrics released. Bots may be identified through names or through heuristics such as rapidly repeated actions. Identifying bots is a non-trivial task that may require some additional research efforts.[12]

[12] For information on how GitHub defines bots, see here: https://docs.github.com/en/graphql/reference/objects#bot

Recommendations

Release metrics, not indicators

A metric is a standard of measurement. Different disciplines have used the term metric in projects on or involving GitHub. For example, the CHAOSS project uses the word metric as “measures of open source project health.” In this report, we define metrics as *data aggregated, across the platform, over several projects.* Metrics can support work in international development, public policy, and economics. On their own, metrics do not imply strength, value, or condition. Examples of metrics include population of a country or the number of commits on a repository.

Indicators, on the other hand, can be used to compare and drive a decision. An indicator may be a composite of several metrics, and can be used to determine relative strength, value, or condition among entities. In the interviews and focus group discussions that informed this paper, the authors heard requests for simple national level metrics as well as complex indicators. As mentioned earlier, in international development in particular, simple metrics are insightful.

Indicators are often created in response to a niche research question and at this stage, it is difficult for GitHub to predict how its data can be effectively incorporated in different indicators. The application of the released metrics, however, will inform this understanding.

Consider requests for project-level metrics

The focus group discussions revealed the need for metrics about specific public code / open source projects. Some of the requests the interviewees and participants raised were:

> “I need to evaluate the sustainability and scalability of digital public goods (DPGs). We need to know how many contributors are in which countries for each DPG.”
>
> ~ Focus group participant 1

> “I am thinking of a map with lines between the location of contributors that interacted (e.g., commented on the same issue). The map of lines shows how distributed a project community is and shows a center of gravity, if there is one.”
>
> ~ Focus group participant 2

> “It would be important to provide how many languages each application supports, to evaluate accessibility.”
>
> “If a repo is very followed/starred/pulled and it is not active for a certain amount of time, have an alert. Our organization would like to see what we could do to support in this arena, particularly as it pertains to projects that could become DPGs (or already are)”.
>
> ~ Focus group participant 3

> “I would like to know that this person’s contribution changed something.”
>
> ~ Focus group participant 4

Releasing repository or organization specific data is not within the purview of this report. GitHub can consider expanding, consolidating, and responding to these requests in a separate project.[13]

Emphasize the consultative process

GitHub has existing goodwill among many of those interviewed that it can leverage to increase the visibility of the metrics should it continue to use consultative processes. The CHAOSS project presents one model for consistent, transparent, and consultative release of metrics “for making open source project health more understandable.”[14] The metrics are officially released once or twice a year following a 30-day comment period. The review and feedback are coordinated on GitHub.

GitHub should also clarify if it cannot release certain metrics. Some of the metrics requested center around user demographics that GitHub does not collect and therefore cannot release. This data minimalism approach at GitHub should be included in communication about metrics releases.

How to share the data

The three personas interviewed had different data needs and resource capacities. They also come from disciplines with different data cultures that shape how they access data. For example, many multilateral organizations, such as the International Telecommunication Union (ITU), share data through Excel or csv files uploaded on their websites. Academic researchers, on the other hand, have long standing

[13] GitHub recently joined the World Bank’s Development Data Partnership: https://datapartnership.org. It has also joined the Industry Data for Society Partnership: https://www.industrydataforsociety.com/announcement.

[14] https://chaoss.community/about/

experience working with APIs, which are especially amenable to programmatic analysis but require more specialized technical knowledge and more stable internet infrastructure.

The final metrics GitHub shares can be available through both data dumps (i.e. csv files) and/or through its APIs. The data could also be hosted on cloud services that allow people to run queries without downloading the data. One of the interviewees suggested that storing the results of some common queries could also make it easier to peruse the metrics and reduce duplicative efforts.

GitHub could choose to share metrics through multiple channels to make the metrics accessible to different stakeholders. All the channels for accessing data should be listed on one landing page under a GitHub domain. This is to distinguish the formal channels (those endorsed and maintained by GitHub). The data should be synchronized across all formal channels.

Sites of relevant organizations, such as UN Global Pulse, ITU, the World Bank, the American Economic Association, and the World Intellectual Property Forum could host GitHub metrics. GitHub can also consider publicizing the metrics through blogs and listserves, such as Data is Plural, and speak about the metrics in academic conferences.

GitHub can learn from the shortcomings of prior data-sharing efforts by other platforms. For example, researchers have found the process behind an API to be a black box (Pfeffer et al., 2022). Additionally, GitHub can follow processes for standardized and transparent documentation, such as those prescribed by Gebru et al. (2021) in “Datasheets for Datasets”. Their process requires answering questions on the motivation behind datasets, and their composition, collection processes, uses, distribution, and maintenance.

Conduct technical research on de-anonymization risks

As previously discussed in the Risk Analysis section, the metrics may unintentionally allow certain individuals, projects, or communities to be identified. While evaluating the risks of data sharing is a socio-political and legal analysis, it can also be a technical analysis. This can help determine the minimums for each metric that ensure specific entities (users, repositories) are not identified.

Enable knowledge sharing across personas

Different personas have varying degrees of experience with GitHub and/or in working with platform data. Practitioners in international development (persona 2) who can drive real-time action using insights from data about GitHub activity have the least familiarity with the platform. GitHub can consider programs for skill and knowledge sharing sessions between the different persona groups. This might also help surface new problem statements to academics (persona 1) that can be answered through metrics released by GitHub.

Conclusion

Releasing aggregate data at this scale would be an unprecedented endeavor for GitHub. It would, for the first time, provide a comprehensive view of how code contributions and software developer activity relate to the SDGs. There are several unknowns, but also immense opportunity in this regard. While far from exhaustive, the recommendations in this document serve as an important first step to help GitHub release metrics that can inform work in international development, public policy, and economics research and programs. This data set can create an opportunity to help establish benchmarks to measure progress of SDGs and possibly reveal inequalities and disparities in technical training, education, and compute power to government services.

As mentioned in the beginning of the report, all possible use cases of these metrics cannot be anticipated ahead of time, and there will likely be several iterations of the metrics. Ultimately, GitHub metrics are an opportunity for the company that hosts most of the world's code to contribute to societal and economic development.

References

1. Agrafiotis, I., Nurse, J. R., Goldsmith, M., Creese, S., & Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity*, *4*(1). https://doi.org/10.1093/cybsec/tyy006

1. Arora, P. (2016). Bottom of the data pyramid: Big data and the global south. *International Journal of Communication*, *10*, 19.

2. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, *106*(51), 21484-21489.

3. Blincoe, K., Sheoran, J., Goggins, S., Petakovic, E., & Damian, D. (2016). Understanding the popular users: Following, affiliation influence and leadership on GitHub. *Information and Software Technology*, *70*, 30-39.

4. Brinkhurst, M. and Crowley, J. (2018). Unlocking Insights from Data: Collaboration with Private Sector creates Cutting-Edge Maps for Disaster Response. NetHope Blog. url: https://nethope.org/2018/09/10/unlockinginsights-from-data-collaboration-with-private-sector-creates-cutting-edge-maps-for-disaster-response/.

5. Buckee, C. O., Balsari, S., Chan, J., Crosas, M., Dominici, F., Gasser, U., ... & Schroeder, A. (2020). Aggregated mobility data could help fight COVID-19. *Science*, *368*(6487), 145-146.

6. Cheng, X., Zhang, Z., Yang, Y., & Zhonghua, Y. (04 2020). Open collaboration between universities and enterprises: a case study on GitHub. *Internet Research*, *ahead-of-print*. doi:10.1108/INTR-01-2019-0013

7. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

8. GH Archive. (n.d.). Research, visualizations and talks. GH Archive Website. Retrieved from https://www.gharchive.org/#resources

9. *GitHut 2.0: A Small Place to Discover Languages in GitHub* (n.d.). Retrieved from https://madnight.github.io/githut/#/pull_requests/2021/4

10. Imtiaz, N., Middleton, J., Chakraborty, J., Robson, N., Bai, G., & Murphy-Hill, E. (2019, May). Investigating the effects of gender bias on GitHub. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)* (pp. 700-711). IEEE.

Section: References

11. Jiang, J., Lo, D., Yang, Y., Li, J., & Zhang, L. (2019). A first look at unfollowing behavior on GitHub. *Information and Software Technology*, *105*, 150-160.
12. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science advances*, *2*(3), e1500779.
13. Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576-11581.
14. Mergel, I. (2015). Open collaboration in the public sector: The case of social coding on GitHub. *Government Information Quarterly*, *32*(4), 464-472.
15. Monitoring, evaluation, research and learning Tech (n.d.) Retrieved from: https://merltech.org/
16. Mombach, T., & Valente, M. T. (2018). GitHub REST API vs GHTorrent vs GitHub Archive: A comparative study.
17. Morozov, E. (2013). To save everything, click here: The folly of technological solutionism. *Public Affairs.*
18. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*. https://doi.org/10.1109/sp.2008.33
19. Ojanperä, S., Graham, M., & Zook, M. (2019). The digital knowledge economy index: mapping content production. *The Journal of Development Studies*, *55*(12), 2626-2643.
20. Perrotta, D., Johnson, S. C., Theile, T., Grow, A., Valk, H. de, & Zagheni, E. (2022). Openness to Migrate Internationally for a Job: Evidence from LinkedIn Data in Europe. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 759-769. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/19332
21. Pfeffer, J., Mooseder, A., Hammer, L., Stritzel, O., & Garcia, D. (2022). This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. *arXiv preprint arXiv:2204.02290*.
22. Pozen, D. E. (2005). The mosaic theory, national security, and the freedom of information act. Yale LJ, 115, 628.
23. Raval, N. (2019). An agenda for decolonizing data science. *Spheres: Journal for Digital Cultures*, (5), 1-6.

Section: References

24. *ROSS Index: the fastest-growing open-source startups, every quarter.* (2022). Retrieved from https://runacap.com/ross-index/
25. Registry: Digital Public Goods Alliance (n.d.). Retrieved from https://digitalpublicgoods.net/registry/
26. Dutta, S., & Lanvin, B. (2021). The network readiness index 2021. Washington: Portulans Institute. Retrieved from https://portulansinstitute.org/introducing-the-network-readiness-index-2021/
27. Taylor, L., & Schroeder, R. (2015). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*, *80*(4), 503-518.
28. Zöller, N., Morgan, J. H., & Schröder, T. (2020). A topology of groups: What GitHub can tell us about online collaboration. *Technological Forecasting and Social Change*, *161*, 120291.