Sustainability of open source science projects

A Data Management Plan created using DMPonline

Creator: Yo Yehudi

Affiliation: University of Manchester

Template: University of Manchester Generic Template

ORCID iD: 0000-0003-2705-1724

Project abstract:

This project investigates the sustainability-related effects of training interventions on open science projects. This training program, branded “Open Life Science” (http://openlifesci.org) which began in early 2020 and was designed collaboratively with the Mozilla Open Leadership initiative, a project which trains project leaders to share their work openly and lead open communities. As part of the program, project leads will be trained in effective community building, (an important part of open science) and open source management techniques for open science projects, over approximately fifteen weeks. Project leads will be able to opt-in to participate in a study tracking their project health from the training sessions and for around a year afterwards. This study will measure project health two ways: Subjectively: Administering questionnaires so project leads can self-assess their project’s health. Objectively, by measuring project health using the Community Health Analytics Open Source Software (CHAOSS) Project (https://chaoss.community/metrics/), which offers both qualitative and quantitative metrics across areas including organisational governance, diversity and inclusion, code quality, licencing, and risk management. Where possible we will script automated measurements, such as the number of stars or contributors a given project’s GitHub repository has, but some metrics are likely to require manual assessment, e.g. assessing quality of documentation or community behaviour guidelines. Follow up measurements will be taken six and twelve months after the training, to see whether project activity is ongoing, and if so, what the project status is.

Last modified: 30-05-2020

Sustainability of open source science projects

Manchester Data Management Outline

- Ethics
- No - only institution involved
- Re-use existing data (please list below)
- Acquire new data

Data will be generated by:

1. Gathering survey data from participants.
2. With participant consent, it will also be gathered from GitHub profiles, using both automated (computationally scripted) and manual data gathering techniques. This will only be data that are already open and clearly shared in the web, and licenced to be re-used.

- University of Manchester Research Data Storage Service (Isilon)
- < 1 TB
- Yes

Some data will be drawn from GitHub (or potentially other open source collaborative platforms). This will all be open data and once gathered, will be stored in Manchester research data storage.

- 21+ years
- Personal information, including signed consent forms

Data gathered will fall under two categories: data which are already open, and data which is more personal.

1. **Open data:** All data gathered from GitHub will be open data that are freely available from the web. This will be data related to contributors to projects, contrbutions to projects, project activity frequency, bugs/issues reported, and licence.
2. **Personal data:** Participants will also complete surveys discussing their expectations for their projects and perceived progress.

The spirit of this study is studying open project behaviours for sustainability. As such, participants will be encouraged to share as much openly as possible, and as mentioned earlier, some of the data will be gathered from sources that are already open anyway. This said, data will be handled in the following ways:

1. **Open data**: Data gathered from GitHub will already be open, and participants will be given the option of letting all metrics data gathered about their project remain open (including their project name) or having it be pseudonymised. Note: it may be potentially possible to identify the project even once pseudonymised, if the data gathered are unique enough. This is likely to be low risk, as these projects are by their nature open and most of the data will be known / retrievable by the public anyway.

2. **Personal data**: A consent form will be part of the survey used to collect survey data about participants expectations for their project, and project success. Data will be stored in the following locations:

- During data gathering phase: in the Manchester-managed survey software, Select Survey
- During data analysis phase: it will be stored on a University of Manchester dedicated laptop with an encrypted solid state disk. The laptop is password protected, has an automatic screen lock timeout after a few minutes, and will not be left unattended whilst unlocked. It is also possible that DropBox for Business (Manchester-managed) will be used for collaboration purposes with other Manchester researchers during this period.
- Long-term data storage will be using the University of Manchester research data storage.

Once data gathering and analysis phases are over, any personal data that was not earmarked to be shared by the participants will be removed from its previous locations (e.g. deleted from the laptop, dropbox, and the survey software, but left on the research data storage).

Personal survey response data will not be pseudonimised and will not be shared openly if participants do not agree to share it, but it may be referred to anonymously in aggregate form - e.g. "52% of projects reported that..." or as an anonymous quote, e.g. "one project stated 'One of our goals is to...'".

- No
- No
- No

- No

Caroline Jay

2020-03-07

Project details

The purpose of this research project is to monitor openly run open source projects over time, and identify possible project aspects that are correlated with ongoing sustainability of these projects.

Information handling minimum controls: http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=30205
Information security classification, ownership and secure information handling SOP http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=29971
Research Data Management Policy http://documents.manchester.ac.uk/display.aspx?DocID=33802
Data Protection policy http://documents.manchester.ac.uk/display.aspx?DocID=14914
Records management policy http://documents.manchester.ac.uk/display.aspx?DocID=14916
Records retention policy http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=6514

Responsibilities and Resources

Yo Yehudi, the student performing this research, will be the primary person performing the steps of the data management.
Caroline Jay, supervisor for the student, will be the Data Custodian.

Question not answered.

Data Collection

Two types of data will be gathered: Survey answers and data gathered from code repositories.
Survey answers: Questionnaires will mostly address project contributors intents and hopes for their projects, with free-text as well as yes/no answers and multi-choice answers. These will be exported as a single comma separated value (csv) file from the survey software. Exported CSVs will be named [date]_questionnaire_[initial/6monthsfollowup/12monthfollowup].csv where the text in square brackets is dynamic depending on the time the questionnaire is being run.
Data gathered from code repositories: there will be two types of data gathered from code repositories - data that are generated by software tools such as grimoirelab will generate json text files, while data gathered by manual assessment of the software repositories will be collected in csv files and generally will be simple yes/no answers or ordinal rankings, with optional textual comments if needed.

Data are likely to number in the megabytes in total and will not represent a serious storage issue.

There will be three phases of data generation, with the possiblity of longer-term follow up phases. The three concrete phases are as follows:
Phase 1 and recruitment:
Potential participants will be invited to participate via email and chat. If insufficient participants are gathered from the target audience, a general call to open source projects will be made via twitter and email groups.
Initial survey: Once recruited, participants will be asked to complete a short survey about their hopes and expectations for their project. Records will generally be kept on a project-level basis, so if a project has more than one interested participant they will be asked to compose their responses as a group, or to nominate an individual to respond on their behalf.
Surveys will ask about information such as the intended scale / longevity of the project, whether it has existing resources online or is nascent, and who the main contributors to this project are.
If the project already has online resources (specifically on GitHub), there will also be data gathering using two methods:
Online data gathering - Automated GitHub measurements:
Tools such as GrimoireLab's Perceval use the GitHub API to gather statistics such as how many contributors a project has and who they are, when issues/bugs are filed and when they are closed (if ever), when code contributions are made and when they are accepted, and so on. Each project will

have the same set of metrics gathered using these tools.

Online data gathering - Manual GitHub measurements:

Some of the data we plan to gather are of a nature that is not easily assess via an automated tool - for example, while an automated tool could check for the presence of a code of conduct file, it is not likely to be sophisticated enough to be able to assess whether there are concrete reporting guidelines for code of conduct violations, so some of the metrics will be gathered by researchers directly checking repositories and reporting the findings.

Phases 2 and 3:

The second phase will be carried out approximately 6 months after the initial phase, and the thrd phase approximately 12 months after the initial phase. Data gathering will be broadly the same on both pases.

Projects will be assessed as active (activity within the last week), active (activity within the last month), active (activity within the last three months), dormant (not active within the last three months, but no closure notice), deprecated (recommends alternatives but still provides updates), or closed (official notice that the project is no longer active).

For all projects, the same range of metrics from the first phase (manual and automated) will be performed again, to monitor changes over time. Projects that were closed permanently in the second phase will not be re-assessed in the third stage.

File formats and storage structure:

All data stored will be in text formats such as csv or json, which are both machine and human readible. Files will be stored on a per-project folder basis, e.g

- projectid/month0/files
  - survey.csv
  - github_pull_requests.json
  - github_issues.json
  - contibutor_stats.json
- projectid/month6/files
  - github_pull_requests.json
  - github_issues.json
  - contributor_stats.json

Projects will be pseudonymised and referred to by a unique id throughout any data we create, but it is likely that data gathered from open GitHub resources may contain the name of the projects and ids of the contributors.

Documentation and Metadata

The data will be accompanied by readme file in the root of the repository. This readme will describe the structure of the folders, and describe clearly how data were generated (e.g. via survey, automated data gathering, or manual curation). In scenarios where analyses were driven by computer code, this will be clearly indicated and there will be a link to the computer source code repository that generated the data. All computer code will be sufficently documented to allow someone unfamiliar with the project to re-run the analysis - possibly in the form of a Jupyter notebook.

Any parts of the data which are published publicly will include clear licence notices making it clear what types of re-use are permissible.

Ethics and Legal Compliance

Personal data: While data will be personally identifiable in the early stages of data gathering, late stages such as aggregated and analysed data results are not personally identifiable. In addition, the risk of inadvertent breaches or de-anonymisation are low - the data gathered are largely open already, and pertain to leading open source projects. There is no reference to sexuality, race, gender, political views, or any other topics that are generally treated as sensitive, unless those data were already shared openly in a participant's GitHub profile.

No survey data will be recorded or shared without the consent of participants. Where data comes from open source software projects and is already in the public domain, people who join the project will be informed about the fact that it may be used for research purposes.

Nevertheless, since this topic involves surveying individuals and small amounts of personal data, the study plans will be reviewed by an ethical review board before the study commences.

Consent: Given that at many times, open source projects will have multiple contributors from across the world, we will ensure that at least one person from the project team has consented to the study and has filled out a complete consent form, but it is possible that individual contributors who come after this consent may not be aware of this. We will provide badge-style notices for project leads to add to their Readme (which functions as a landing page), ensuring that they are displayed on the project page before we run the GitHub data collection process.

Data will be shared openly and re-usable under an open licence (probably CC0 Public domain, to facilitate re-use), and people will be encouraged to credit their source if they re-use the data.

Source code used to generate any results will be open source and licenced under a permissive non-copyleft licence such as MIT.

Storage and backup

Data will be stored on the macbook used to collect it and backed up in Research Data Storage.

Data will be stored on a macbook with an encrypted hard disk. Access to the machine is password-protected and the machine is always locked when unattended, and locks on sleep/closed lid. There will be only one user with access to the Macbook.
Transfer of any personally identifiable data to the Manchester Research Data Storage facility will be performed via the Manchester VPN.

Selection and Preservation

This study aims to share all data for re-use where possible, with the exception of data that must not be shared for personal privacy reasons.
All other data will be preserved for at least 5 years after publication in line with the University of Manchester's policies, but ideally indefinitely.
Data which was gathered from open data sources such as GitHub, and data where the participant has expressly agreed to share openly will be shared and deposited in a data repository.

As above - data which has been aggregated to produce publishable results and data where the participant has expressly agreed to share openly will be shared and deposited in a data repository.

Data Sharing

Data which has been aggregated to produce publishable results and data where the participant has expressly agreed to share openly will be shared and deposited in a data repository such as Zenodo.

Data - in the form of survey responses - will not be shared unless the participant has expressly consented that it be shared.
Data will only be drawn from GitHub (where it is already displayed openly) if there is a licence on the repository that permits re-use.