Mining Software Repositories Hackathon: World of Code

Audris Mockus
EECS
University of Tennessee
Knoxville, USA
audris@utk.edu

James Herbsleb
School of Software Engineering
Carnegie Mellon University
Pittsburgh, USA
email address or ORCID

Alexander Nolte
Computer Science
University of Tartu
Tartu, Estonia
alexander.nolte@ut.ee

***Index Terms*—Mining Software Repositories, Hackathon**

I. Call for Participation

The proliferation of open source software (OSS) development has created novel and massive software supply chains and ecosystems that require non-traditional approaches of software development. Even the approaches that fared well in the earlier stages of open source development are challenged by the shear scale of the present open source ecosystem, the complexity of dependencies among projects, and the lack of effective means of establishing trust essential for frictionless collaboration - a cornerstone of OSS.

The purpose of the World of Code (WoC) hackathon is to explore solutions that either apply at a global scale or require measurement approaches done at that scale. Examples for these are any measurements such as complete OSS activities of developers, complete downstream dependencies of a project, or the provenance of a source code file.

Please join if you are concerned about the continued health of open source software and would like to make a difference! The teams selected by the PC (see more detail below) will have the opportunity to publish the description of their work at MSR'2021. Previous WoC hackathon resulted in four publications at MST'2020.

Any topics related to doing research, building tools, or improving infrastructure that supports global OSS development are within the scope of the hackathon. For example,

- Applications that support finding suitable code, people, projects, or bugs.
- Applications that increase transparency by making it easier to become a contributor or that helps maintainers zero in on most relevant contributions.
- Applications that increase understanding of software supply chains and ecosystem: how and why they function and how to manage risks.
- Any infrastructure work that does data fusion or data quality improvements, such as leveraging all open source data sources in WoC resource and beyond.
- Approaches to better collect data increase the coverage or encourage outside contributions.

II. Key Dates

The CFP will be released by mid-September.

The project ideas from participants will be collected until mid-October.

The date for hackathon itself will be arranged by conducting the survey of the project idea submitters but will be conducted before the end of 2020.

III. The Process

An online hackathon for researchers who would like to team up with others on crucial problems for supporting Open Source Software. Organizers will provide training for World of Code resource that supports global measurement of OSS in the form of a webinar prior to the event.

The event will provide activities typical of the in-person hackathon virtually. For example, defining research questions, forming teams, scoping problems. This will be done while also providing advice on the best ways to conduct data processing and improve performance.

Organizers will provide support in the form of mentors that can help with technical issues. The hackathon will also provide the opportunity for participants to work with world-class researchers on relevant problems and research questions.

The preparation will involve team forming activities and a set of online tutorials during which we will walk MSR Hackathon participants through key functionality.

A dedicated (issue tracker) to answer questions and solve issues for the participants of MSR Hackathon will also be available. Slack and Zoom will be suggested as the main means of communication during the hackathon between teams, mentors and organizers. A dedicated Zoom room will be provided for each team during the entire duration of the event.

The results of the hackathon will be evaluated by the Hackathon Evaluation Committee based on the originality of the idea, potential impact of the proposed solution, and the sophistication of the artifacts produced during the hackathon.

The winning teams will have an option to present their results at a special session of Mining Software Repositories conference 2021.

IV. Proposed Hackathon Evaluation Committee

Name	Affiliation
Chris Bird	Microsoft
Kelly Blincoe	University of Aucland
Premkumar Devanbu	University of California
Roberto Di Cosmo	Software Heritage
Anna Filippova	GitHub
Daniel German	University of Victoria
Georgios Gousios	TU Delft
Ahmed E. Hassan	Queens University
Abram Hindle	University of Alberta
Rajan, Hridesh	Iowa State University
Tim Menzies	North Carolina State University
Emmerson Murphy-Hill	Google
Mei Nagggapan	University of Waterloo
Gregorio Robles	Universidad Rey Juan Carlos
Diomidis Spinellis	Athens University of Economics and Business
Peggy Storey	University of Victoria
Lin Tan	Purdue University
Laurie Williams	North Carolina State University
Bogdan Vasilescu	Carnegie Mellon University
Minghui Zhou	Peking University

V. World of Code Resource for Hackathon

WoC version R represents 123M git repositories from GitHub, GitLab, BitBucket, etc. based on updates/new repositories identified on Mar 7, 2020, and retrieved by Mar 28. It has 2B git commits (42M authors), 8B blobs, and 8.3B trees. The content of these objects is augmented via cross-referencing (a graph) for immediate access to all object references associated with any specific object. For example, all commits that have created a specific blob, all repositories where a specific blob or a specific commit resides in, all commits for a specified author ID, the child commits of a specific commit, and other maps that are impossible to compute without complete data.

Developers often have multiple commit author IDs. WoC solves this problem via a map that associates each author ID with imputed developer identity [1]. Some of these are actually bots [2].

With many repositories being clones or forks, WoC for each repo also provides an estimated project identity for each repository [3]. This results in 69M independent projects.

Time-based selection of commits and selection of projects and authors based on the characteristics of files and activities associated with them is also provided.

WoC pre-calculates basic summaries for each blob by running ctags and stores the results as well as diffs for all 2B commits. Finally, WoC provides blob to import/include/use statements parsed from blobs in 17 programming languages.

a) Access.: Due to its large size, WoC provides ways to easily access, sample, and analyze the data based on all the cross-referencing capabilities. Shell, Python, and Perl API can be use on WoC servers via ssh (high performance) and REST APIs for web access. Users need to fill in a short web form for ssh access (see Appendix II).

b) Participant needs.: Participants who are comfortable using terminal would benefit from high performance of ssh access where two simple shell commands showCnt and getValues can be composed with regular shell commands into an arbitrarily complex analysis as described in the tutorial. Participants who disdain terminal can use Python in Jupyter notebooks on WoC servers [4] or run Jupyter notebooks on participants' computers and access WoC via REST APIs [5]. The primary focus of WoC is to provide cross-referencing and completeness of the coverage of the git objects from public repositories. Participants will be responsible for integration of WoC data with other types of data they may need: e.g., issue metadata from GHtorrent, code parsing, statistical modeling, and and other types of data and analysis (See Appendix IV for an example workflow).

c) The research questions.: The purpose of WoC is to enable answering questions that can not be answered without the completeness and cross-referencing of public git data. Research on networks of developers, code, and API's would be benefit the most. For example, questions that require the measurement of activity of the entire OSS at a particular slice of time, identifying all commits/repositories/blobs/developers associated with a specific developer/blob/API, finding all child commits of any commit, or finding all forks/clones of a repository. Simple tasks like finding all commits/blobs associated with a repository are supported but do not require WoC, as "git clone" would suffice. WoC can be used for representative sampling of the developers and projects by developer and project network properties [6]. Such complete picture allows to go past MSR analyses limited by data filtered on the set of non-representative projects and, therefore, unable to capture the full scope of developer activity, code reuse, or API usage. Mining questions might involve improving quality of WoC data, its performance, ways to link to other datasets. WoC use examples are in [7] and in publications in Appendix III. The evolution of open source software, developer and project ecosystems, past and current trends, counting and summarizing developers, projects, the source code, and the dynamics of developer and code movements across entire OSS ecosystem are examples of research questions that could not be answered without WoC. Answers to these questions have major practical implications for developers and industry alike: will a library becomes obsolete? Do we have any open source code?

d) Sample data.: Since sampling large graphs is difficult, we provide a Jupyter notebook [8] that investigates blobs/projects/authors associated with a Common Vulnerability and Exposure (CVE).

Appendix I: The Access Form

MSR Hackathon registration

Any identifying information provided here will remain confidential. Your participation in this event is entirely voluntary.

* Required I am age 18 or older, I have read and understand the above information and I wish to participate in the upcoming hackathon.

* Yes/No

Demographics and event organization

First name *

Last name *

Email address *

Affiliation *

Gender *

Female/Male/Non-binary/Prefer not to say/Other

What is highest level of formal education that you have completed until now?

High school diploma or GED/Some college Associate and/or bachelor's degree/Bachelor's degree/Professional degree/Master's degree Doctorate/Prefer not to say

What is your current occupation? (e.g. data scientist, PhD student, post-doc, faculty)

Access to World of Code

This information is required to add you to the corresponding organizations on GitHub and create your login access to the World of Code systems. GitHub handle

Preferred login name for World of Code *

Public ssh key * Please refer to https://help.github.com/en/articles/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent for information on how to create it if you do not have one already.

On a scale from 1 to 10, how do you estimate your current experience using World of Code?

World of Code hackathon

How many hackathons have you participated in the past? If you cannot recall the exact number please provide a rough estimate.

To what extent was your decision to participate in the upcoming World of Code hackathon motivated by:

What if anything at all did you do to prepare for the upcoming World of Code hackathon?

Do you have an idea you would like to work on during the hackathon? If yes, please describe it briefly below.

REFERENCES

[1] T. Fry, T. Dey, A. Karnauch, and A. Mockus, "A dataset and an approach for identity resolution of 38 million author ids extracted from 2b git commits," in *IEEE Working Conference on Mining Software Repositories: Data Showcase*, May 2020. [Online]. Available: https://arxiv.org/abs/2003.08349

[2] T. Dey, S. Mousavi, E. Ponce, T. Fry, B. Vasilescu, A. Filippova, and A. Mockus, "Detecting and characterizing bots that commit code," in *IEEE Working Conference on Mining Software Repositories*, May 2020. [Online]. Available: https://arxiv.org/abs/2003.03172

[3] A. Mockus, D. Spinellis, Z. Kotti, and G. J. Dusing, "A complete set of related git repositories identified via community detection approaches based on shared commits," in *IEEE Working Conference on Mining Software Repositories: Data Showcase*, May 2020. [Online]. Available: https://arxiv.org/abs/2002.02707

[4] [Online]. Available: https://github.com/woc-hack/msr-challenge/PYJupyter.ipynb

[5] [Online]. Available: https://github.com/woc-hack/msr-challenge/blob/master/RESTJupyter.ipynb

[6] A. Tutko, A. Henley, and A. Mockus, "More effective software repository mining," in *http://arxiv.org/abs/2008.03439*.

[7] Y. Ma, C. Bogart, S. Amreen, R. Zaretzki, and A. Mockus, "World of code: An infrastructure for mining the universe of open source vcs data," in *IEEE Working Conference on Mining Software Repositories*, May 26 2019. [Online]. Available: papers/WoC.pdf

[8] [Online]. Available: https://github.com/woc-hack/msr-challenge/CVEJupyter.ipynb