2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion) | 978-1-6654-1219-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICSE-Companion52605.2021.00111

An Open Dataset for Onboarding new Contributors– Empirical Study of OpenStack Ecosystem

Armstrong Foundjem[†], Ellis E. Eghan[‡], Bram Adams[†]
MCIS Laboratory — ([†]Queen's University, [‡]Polytechnique Montréal, Canada)
{a.foundjem, bram.adams}@queensu.ca, ellis.eghan@polymtl.ca

***Abstract*—This dataset provides the qualitative and quantitative data of our mixed-method empirical study of onboarding in the OpenStack software ecosystem (SECO). First, we carried out a SECO-level participant observation study of 72 new contributors during a 2-day OpenStack onboarding (in-person) event yielding a rich set of qualitative data; 14 files amount to 60% of the entire dataset originating from a participant observation study. Second, we quantitatively validated the extent to which SECOs achieve benefits such as diversity, productivity, and quality by mining 1281 contributors' code changes, reviews, and issues with(out) OpenStack onboarding experience. Our quantitative dataset includes nine files, which is about 40% of the entire dataset, and we obtained these files by mining new contributors' codebase activities from four OpenStack repositories. Besides, we make available the scripts that e used to extract and analyze this dataset. By providing this data, we are claiming the "Available Badge," and our data are online on a public archived repository at Zenodo: DOI: 10.5281/zenodo.4457683**

***Index Terms*—Available, open data, open science, replication, verifiable, transparency**

I. General Description and Background

In the interest of open science [1], we make the data of our mixed-method research available to researchers, SECOs, and Companies [2].

A. SECO-level Onboarding Event —Qualitative Study

A SECO's onboarding program follows a two-phase "continuous" process. A top-level training (in-person two-day event) followed by a project-specific training (remote event); usually, one-one sessions between mentor-mentee(s) until the mentee makes their first acceptable changes in the codebase. The provided qualitative data originates from a 2-day SECO-level observational study onboarding event on 72 OpenStack new contributors willing to join the ecosystem. Figure 1 shows the seating configuration for both observed participants and mentors on each table (T1,..., T12). We also randomly observed participants as they perform tasks using the think-aloud protocol [2]. We used four high-quality professional audio-visual equipment (C#1,..., C#4) to record the entire events, which we later transcribed and analyzed.

1) Observation Dataset: Our qualitative dataset in the 1.codebook folder contains five files: 1.day1-2-observation.pdf and 2.technical-activities-onboarding.pdf, which are the transcribed files of the 2-day observation study. The 3.onboarding-emerging-codes-irr.pdf file contains themes and categories from the transcribed text that both coders identified during the inductive coding phase. First, each coder independently coded the transcribed text, and then both coders compared the themes and to which category these themes belong. Usually, this inductive coding has several rounds of a negotiated agreement. Coders/reviewers have to argue if a particular theme or category should/shouldn't be under a new/existing arrangement. This process also involves the computation of inter-rater reliability (IRR).

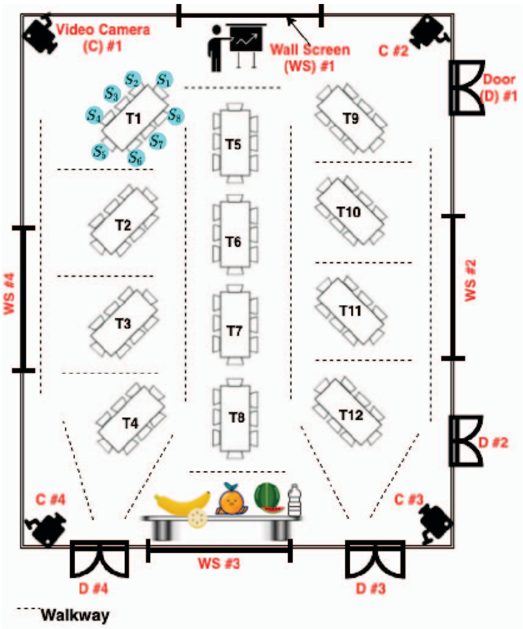

Figure 1. Participants' seats are configured in clusters during a typical SECO top-level event. Tables ($T_1, \ldots, T_{12}$) each had eight seats ($S_1, \ldots, S_8$). Each side of the hall is equipped with four screens ($WS_1, \ldots, WS_4$) alongside four audio-visual cameras ($C_1, \ldots C_4$ to record the entire event.

Next, the 4.think-aloud-protocol.pdf file was generated when the principal observer OB1 randomly asked participants to explain a task they are performing. Last, the codebook 0.Codebook-with-Examples.pdf contains the outcome of the qualitative coding activity. This file has three columns. The first column shows code that originated from the inductive phase. The second column provides a description and the rationale of each code, and in the third column, we give concrete examples of how we extracted each code from the transcribed text contextually.

2) The Codebook Building Process: To build our codebook, both coders of the qualitative data did inductive

978-1-6654-1219-3/21/$31.00 ©2021 IEEE
DOI 10.1109/ICSE-Companion52605.2021.00111

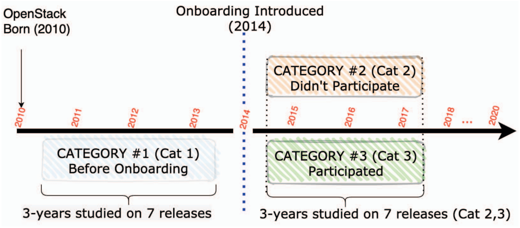

Figure 2. Timeline of stratified categories used in our quantitative study. Cat-1 is our control group, while Cat-2 and Cat-3 are the experimental groups. Each group uses data of seven OpenStack releases (42 months).

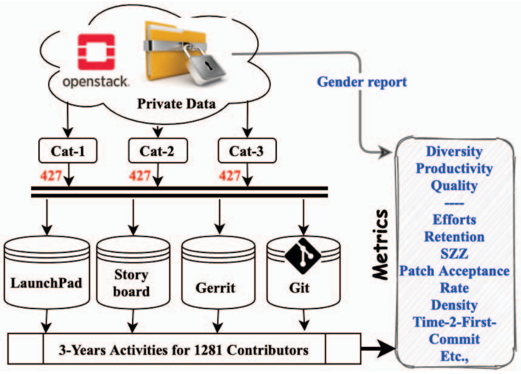

Figure 3. The quantitative dataset is based on Table1-Metrics.pdf. Except for gender reports (Confidential), all other data are extracted from the OpenStack repositories, such as Gerrit, Issues trackers (LaunchPad and Storyboard), and Git-based OpenDev system.

(1.qualitatve...csv) and deductive coding containing four rounds of inter-rater reliability (IRR). We capture these rounds in folder 2.irr-iterations, 3.irr., ..., 6.irr, and report the outcome of the IRR in the 2.irr-onboarding.csv file. The IRR result is a hierarchical-structure having three levels of depth, H1, H2, and H3. Also, both coder one/two use several labels to indicate a (dis)agreement of a code in the text. For example, TESTIMONY1r1, ..., OVERVIEW2r2, with entering of either a 1/0 in each cell.

3) Affinity Diagram: After several iterations of comparing common themes, higher level themes started emerging to forming an high-level abstraction of the emerged code, we categorized themes in a hierarchical structure, which we represented in an Affinity diagram (3.affinity-diagrams-iterations). We did three iterations of negotiated agreement, arranging the codes in different configurations to obtain the final design 3.affinity-final-iteration.pdf. This affinity diagram becomes the qualitative data outcome of the SECO onboarding event: catalog of Teaching content, Challenges, and Benefits of onboarding. Thus, we select the most prominent activities based on how much time participants/mentors spent.

B. Study on Contributors' productivity, diversity and Quality

Our quantitative analysis compares three categories of contributors Cat-1 vs. Cat-2, and Cat-2 vs. Cat-3 (see Figure 2) in terms of productivity, diversity and quality. If no difference exists between Cat-1 vs. Cat-2, then we assume that any difference between Cat-2 vs. Cat-3 will be correlated to onboarding effects. Our scripts are available in the 5.scripts folder, and figure 3 shows the repositories that we mined. We used in-house/open-source tools (https://opendev.org/x/stackalytics.git, and https://github.com/chaoss/grimoirelab-perceval.git) to extract and analyze contributors' data from OpenStack repositories, such as issues tracker, Gerrit, Git (OpenDev). The main script to analyze data for various metrics during new contributors mentoring activities is (11.onboarding.html), and the statistical analysis is (12.Statistical-testing.html). Each of these files is an *.html* copy of the Jupyter notebook. Our R-scripts list their dependencies at the first lines of each file. Users can run our script by specifying the dataset's path in their system and obtaining desired results. For example, users can create the same path in 13.survival-analysis.R to analyze the time-to-event (retention) for each new OpenStack contributor in Cat-1, Cat-2, and Cat-3. To measure the impact of onboarding on contributors, we run the python script (12.Statistical-testing.html).

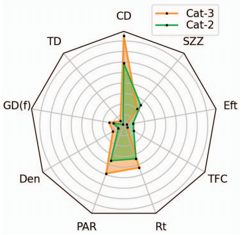

Figure 4. Findings suggest that onboarding (Cat-3) has significant differences and improvements over Cat-2 on Bug-inducing-commits (SZZ), Effort (Eft), Time to first commit (TFC), Retention (Rt), Patch Acceptance Rate (PAR), Density (Den), Diversity: Gender (GD(f)), Technical (TD), and Corporate Diversity (CD).

II. Usage and Summary of Artifacts

To use/replicate our research, we have made available our dataset and scripts in sequential order: *1.codebook, 2., ..., 5.scripts.* This data comes from both the qualitative and quantitative research methods. In particular, researchers can decide to run only the quantitative or the qualitative script/data. We used python 3.8.5 with the dependencies mentioned in *requirement.txt*. R version 4.0.3.

In sum, our "*21.radar.py*" script summarizes the metrics that we used in the quantitative analysis comparing Cat-2 vs. Cat-3 contributors and shows a statistically significant difference in all the studied metrics. With a chart such as shown in Figure 4, users can compare two or more groups for a studied phenomenon. We use a log-log scale to normalize the data for both groups.

References

[1] A. Rahman, M. R. Rahman, C. Parnin, and L. Williams, "Security smells in ansible and chef scripts: A replication study," *ACM Trans. Softw. Eng. Methodol.*, vol. 30, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3408897

[2] conferencepapers, "conferencepapers/icse_2021: Final release icse21," Jan. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4457683