Human Computation and Crowdsourcing: Works in Progress Abstracts:
An Adjunct to the Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing

The Rise of Curation on GitHub

Yu Wu[1], Jessica Kropcznyski[2], Raquel Prates[3], John M. Carroll[4]

[1,2,4] College of Information Sciences and Technology, The Pennsylvania State University

[3] Computer Science, Federal University of Minas Gerais

yuw132@ist.psu.edu, jnk14@ist.psu.edu, rprates@dcc.ufmg.br, jcarroll@ist.psu.edu

Abstract

Recently, curation practices start to develop in GitHub, where developers systematically put efforts to select, evaluate, and organize existing artifacts for the purposes of preservation and future use in software development. Curation practices in social media sites, such as Twitter and Pinterest, have been investigated, raising questions about the nature of collaborative curation in a professional/product-oriented site. In this study, we identify and characterize curation projects hosted on GitHub, and compare curation projects with software projects to study how this practice takes place and how it is different from the original use of GitHub. We find that curation has emerged as a highly popular category of GitHub project, which is directed to learning and professional development, and curation practice leverages collaborative tools and practices native to GitHub. Although curation projects and software projects use the same set of activities for development, they are different from each other in terms of the quantity of each type of activity performed by developers.

Motivation

Software developers increasingly utilize GitHub for coordinating work (Marlow et al. 2013). By the end of 2013, GitHub had hosted more than 10 million repositories and attracted millions of developers for collaborative software development (Doll 2013). Recently, developers have begun to create GitHub repositories to index resources related to software development itself, and many such repositories became trending on the site. This category of activity, which is entirely different from development activity, is known as curation. Curation is the manual effort of identifying, organizing, evaluating public resources (Duh et al. 2012). As GitHub is designed for collaborative software development, this phenomenon poses interesting questions of how such practice emerges and becomes trending, and how the community of software developers is doing with it as compared to the original purpose of GitHub.

Although past literature reports that software developers use GitHub for a variety of purposes (Wu et al. 2014), most studies focus on software engineering practices, such as code contribution and bug fixing (Marlow et al. 2013). The characteristics of curation practice on this social production site, to the best of our knowledge, have not yet been reported. This paper intends to address this gap and the following research questions are addressed: ***RQ1****: How is curation activity emerging in GitHub*? And ***RQ2****: How does curation practice take place differently as comparing to software collaboration on GitHub?* Our investigation tries to understand the intention and content being curated on a social coding site and specifically, how such practice is different from the original intention of the platform. Our work extends curation literature in the context of software developers' community.

Data Collection

We identified the top 49 curation projects and 1,384 software projects from the dataset collected from GitHub Archive[1], then we performed a content analysis on the 49 projects to answer RQ1. The unit of analysis was each individual curation project. Then, we aggregated the activity log of each of the 49 curation projects and 1,384 software projects respectively, and applied a quantitative method to compare them. Specifically, for each type of activity, we analyzed whether the number of the activities for curation projects is different from the number of activities for software projects. The results are presented below.

Results

The characteristics of curation projects

We applied content analysis to investigate the characteristics of the curation projects, including purposes for creation, contents being curated, and owners. Out of 49 projects, 7 claim in their project description that the project was created for learning, and 23 express the intention to maintain the high quality of content. For the 49 curation projects, two (4.1%) were owned by GitHub organizations and the other 47 were owned by individuals. Correspond-

[1] https://www.githubarchive.org/

ingly, for the 1,384 software projects in our sample, 522 were owned by an organization (37.7%). Curation projects on GitHub try to index the resources of high quality both inside and outside GitHub in a central place for informational, learning, and professional development purposes. Nearly eighty percent of owners had less than 100 followers prior to their creation of curation projects, which indicates that most owners were not "coding rock stars" or community leaders, differently from the results in Matthews et al. (2014) for enterprise environments.

Curation Projects vs. Software Projects

Of the top 1,433 repositories that have more than 500 stars (most popular), 49 are curation projects. The top 3 repositories are all curation projects, and 6 out of the top 20 repositories are curation projects.

Discussion

Given the large volume of resources inside and outside GitHub for software development (Wu et al. 2015), software developers' community is in need of efforts to index high quality ones, and curation practice on GitHub responds to such needs. Software developers appropriate GitHub to systematically organize resources of high quality for learning and professional development purposes. The owner of curation projects on GitHub are usually normal individuals as comparing to enterprise context, which tends to have a small leadership team that creates and maintains curation projects (Matthews et al. 2014). Given the highly popular nature of curation and its nascent phase, it is likely that more software developers will take on different curation tasks shortly.

Type of Event	Curation (N=49)			Software (N=1384)			*df*	*t*
	Mean	*Mean (log)*	*SD (log)*	*Mean*	*Mean (log)*	*SD (log)*		
Create	1.84	0.80	0.62	16.97	1.85	1.34	65.22	11.01**
Delete	0.65	0.21	0.57	9.40	1.04	1.22	64.91	9.41**
Fork	465.69	5.45	1.08	161.81	4.51	1.12	51.73	-6.04**
Issue Comment	105.10	4.02	1.28	406.91	4.56	1.86	55.48	2.85*
Issues	27.16	2.62	1.27	140.64	3.77	1.65	53.86	6.12**
Pull Request	151.69	4.29	1.28	91.52	3.22	1.59	53.38	-5.76**
Push	405.82	4.57	1.07	165.36	3.85	1.68	56.80	-4.52**

* indicates the result is significant at $p < 0.01$, ** indicates the result is significant at $p < 0.001$
(*log*) indicates the statistics are calculated on log-transformed value.

Table 1: Independent sample t-tests that compare the log-transformed mean of the number of seven types of activities between curation projects and software projects.

Curation projects received a statistically significant higher number of stars (M=7.51, SD=0.99) than software projects (M=6.98, SD=0.65), $t(49.48)=-3.70$, $p < 0.001$. Figure 3 shows that the average number of stars (not log-transformed) received by curation projects is a lot higher than software projects.

We also compared the curation projects and software projects in terms of participation of different types of user activities (Table 1). Curation projects are statically higher in *Fork*, *Push*, and *Pull Request* events, while lower in *Create*, *Delete*, *Issue Comment*, and *Issues* event. Lower frequencies on *Create* and *Delete* events for curation projects imply that they are less likely to have changes in project structures in comparison to software projects. The lower frequencies of *Issues* and *Issue Comment* events for curation projects indicates curation projects are less active on issue trackers, which possibly means that they have fewer bug reports and feature requests. Higher frequency of *Push* events shows the active development of curation projects. Higher frequency of *Pull Request* and *Fork* events indicates that curation projects received more contributions from others.

References

Dabbish, L., C. Stuart, J. Tsay, and J. Herbsleb. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proc. CSCW '12*, 1277-1286.

Doll, B. 10 Million Repositories. 2013; Retrieved on Sept. 2, 2014 from: https://github.com/blog/1724-10-million-repositories.

Duh, K., T. Hirao, A. Kimura, K. Ishiguro, T. Iwata, and C.-M.A. Yeung. 2012. Creating Stories: Social Curation of Twitter Messages. In Proc. ICWSM '12.

Marlow, J., L. Dabbish, and J. Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in Github. In *Proc. CSCW '13*, 117-128.

Matthews, T., S. Whittaker, H. Badenes, and B. Smith. 2014. Beyond end user content to collaborative knowledge mapping: interrelations among community social tools. In *Proc. CSCW '14*, 900-910.

Wu, Y., Kropczynski, J., Shih, P. C. and Carroll, J. M. 2014. Exploring the ecosystem of software developers on GitHub and other platforms. In *CSCW '14 Companion*, 265-268.

Wu, Y., Kropczynski, J., & Carroll, J. M. 2015. Making Big Data Transparent to the Software Developers' Community. *Big Data: Algorithms, Analytics, and Applications*, 175-190.