Research Ethics for Studying Open Source Projects

Christopher Oezbek
Freie Universität Berlin
Institut für Informatik
Takustr. 9, 14195 Berlin, Germany
oezbek@inf.fu-berlin.de

Abstract

The public visibility of Free and Open Source Software development has sparked interest in the research communities of business, social and computer sciences to use the projects as research subjects. This article tries to open a discussion about the implications of this interest, whether the Free and Open Source communities appreciate being under "surveillance" and how we can deal with the ethical problems related to the human subject research.

1 Introduction

With the rising popularity of Free Software and Open Source, numerous researchers have started to explore their field of study within this new development paradigm for software. The main advantage for the researcher is that the communication of the project and the resulting piece of software is available publicly and does not require contacting companies and negotiating NDAs. The bulk of empirical work in this vein uses one of the four following research methods:

- Data Mining is the automated processing of public available metadata offered by the projects' infrastructure such as version control, bug tracking and email lists and the generation of insights from such data.
- Qualitative data analysis is the manual interpretation of unstructured, non-numeric data such as emails. Typically such analysis uses methods from the social sciences such as Grounded Theory [5] for building theoretical models about some aspect of Open Source development.
- Survey research has been used widely by the different research community to better understand the less technical aspects such as the motivation of participation in the Open Source world (for instance [8, 7, 10]). Unfortunately the large number of surveys conducted with Open Source communities have caused some contact mails to be perceived as spam [4].
- Lastly Action Research [1, 2] is a research method in which the researcher and the project collaborate on solving problems jointly [9] or in which the researcher engages as a project participant [12]. In contrast to the other three kinds, action research is not primarily descriptive regarding the projects, but rather tries to come up with novel solutions or evaluate their applicability.

Two dimensions along which these methods can be contrasted are the need of project participants to be active for the methods to work and the benefit that can be achieved for the projects.

With data mining and qualitative data analysis the project participants only seldom need to become active, in fact in many cases they will not be even aware that somebody is taking their publicly available data and analyzing it. Surveys on the other hand need a large sample of members of many different projects to be representative, which leads to many projects being contacted in an impersonal manner. Action research needs highly active project members who negotiate research goals and then pursue them together with the researcher.

Regarding benefit, action research holds the greatest promise, since project and researcher work explicitly towards generating such benefit for the project. Qualitative data analysis and data mining can give participants insights into their own project or personal performance and thus can be beneficial as the basis for improvements. Surveys generally only answer questions about Open Source development at large and thus do not provide project specific insights in most cases.

Concluding, all four kinds involve humans as participants, collaborators or data sources and thus need to be reviewed regarding its ethical aspects and whether they constitute human subject research [13].

2 Human subject research

Traditionally human subject research underlies strict regulation by university advisory boards to ensure that subjects are protected from possible harm. Such protection is usually based on general principles like *autonomy*, *benefice* and *justice* as set forward in the Belmont report [11] or the declaration of Helsinki [14]. While benefice and justice relate to maximizing benefit and minimizing risk without disproportionally affecting a certain group of people, autonomy is meant to ensure that subjects remain in control of what is happening to them [3]. Typically these principles lead to regulations that require the researcher to inform the participants in a study about the goal, possible harm and benefits of the study and gathering their consent for participating voluntarily in the research, anonymizing gathered data to prevent harm caused by public disclosure of private information and employing a review board for human subject research. For survey research these principles can be relatively easily applied. Subjects are informed beforehand about the survey and the amount of time participation will take and their opting-in to respond to the question can be seen as an informed consent. Furthermore, most surveys are automatically anonymous or can be easily anonymized by aggregating data for presentation. Action research by its nature involves detailed discussion with the project about possibilities, risks and chances of the collaboration. It might be argued that Open Source projects in such a discussion are highly autonomous and due to their decision structure basically immune to “bad” outside influence. Problems thus usually only arise if the scientist (1) does not want to reveal his or her identity as a researcher, essentially deceiving the project, or (2) has a preset agenda and steers the discussion towards certain problems or solutions, thus diminishing the autonomy of the project.

Data mining and qualitative data analysis have bigger problems [6]. First and foremost it might be impossible to maintain confidentiality. Even when anonymizing results, it often remains easy to uncover the identity of projects or project participants by simple searches. Verbatim quotations or printing graphs showing number of commits per project members thus are impossible if the anonymity of subjects is of interest.

Second, the implications of publishing results about individual projects or project members are unclear. For instance, negative results in a comparison of quality attributes with other similar projects might make developers or users abandon the project or worse have implications for participants in those projects with regards to their ability to secure a job position. Even positive results in studies might cause harm, as the resulting publicity could attract large crowds of users not interested in contributing. Exactly these unclear implications of publishing results have caused principles like anonymity to be put in place. The researcher will just not be able to foresee all consequences of the data being publicized.

Thirdly, since communities change over time, it might not be possible to gather consent from participants who have left the community since. This problem becomes especially grave if we consider gathering consent from the community as a whole, because we have analyzed their aggregate data statistically.

3 Conclusion

This paper has highlighted some problems with doing empirical research with Open Source communities, especially when using publicly available data without consideration. To put research on a solid ethical base, a discussion between the scientific and Open Source community is necessary to clarify the following questions:

- Should the Open Source world demand that all researchers register their studies publicly so that any possible abuse can be prevented?

- What should the stance towards negative results be? Can negative comments on individual persons be published?

- What harm can plausibly be expected and thus should be guarded against?

References

[1] David E. Avison, Francis Lau, Michael D. Myers, and Peter Axel Nielsen. Action research. *Commun. ACM*, 42(1):94–97, 1999.

[2] Richard L. Baskerville. Investigating information systems with action research. *Commun. AIS*, 2(3es):4, 1999.

[3] Joan Cassell. Ethical principles for conducting fieldwork. *American Anthropologist*, 82(1):28–41, March 1980.

[4] Hyunyi Cho and Robert LaRose. Privacy Issues in Internet Surveys. *Social Science Computer Review*, 17(4):421–434, 1999.

[5] Juliet M. Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, March 1990.

[6] Gunther Eysenbach and James E Till. Ethical issues in qualitative research on internet communities. *British Medical Journal*, 323(7321):1103, 2001.

[7] Rishab Aiyer Ghosh, Bernhard Krieger, Ruediger Glott, Gregorio Robles, and Thorsten Wichmann. Free/Libre and Open Source Software: Survey and Study – FLOSS. Final Report, International Institute of Infonomics University of Maastricht, The Netherlands; Berlecon Research GmbH Berlin, Germany, June 2002.

[8] Alexander Hars and Shaosong Ou. Working for free? - motivations of participating in open source projects. In *The 34th Hawaii International Conference on System Sciences*, 2001.

[9] Letizia Jaccheri and Thomas Østerlie. Open source software: a source of possibilities for software engineering education and empirical software engineering. In *Proceedings of the 29th International Conference on Software Engineering Workshops (ICSEW '07)*, Washington, DC, USA, 2007. IEEE Computer Society.

[10] Karim R. Lakhani and Robert G. Wolf. Why hackers do what they do: Understanding motivation and effort in Free/Open Source Software projects. In Joseph Feller, Brian Fitzgerald, Scott A. Hissam, and Karim R. Lakhani, editors, *Perspectives on Free and Open Source Software*, pages 3–22. The MIT Press Ltd., Cambridge, MA, July 2005.

[11] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report: Ethical principles and guidelines for the protection of human subjects of research, 1979. `http://en.wikisource.org/wiki/Belmont_Report`, visited 2008-01-27.

[12] Christopher Oezbek and Lutz Prechelt. On understanding how to introduce an innovation to an Open Source project. In *Proceedings of the 29th International Conference on Software Engineering Workshops (ICSEW '07)*, Washington, DC, USA, 2007. IEEE Computer Society.

[13] Joseph B. Walther. Research ethics in internet-enabled research: Human subjects issues and methodological myopia. *Ethics and Inf. Tech.*, 4(3):205–216, 2002.

[14] World Medical Association. Declaration of helsinki: Ethical principles for medical research involving human subjects, 2000. `http://www.wma.net/e/policy/b3.htm`, visited 2008-01-27.