

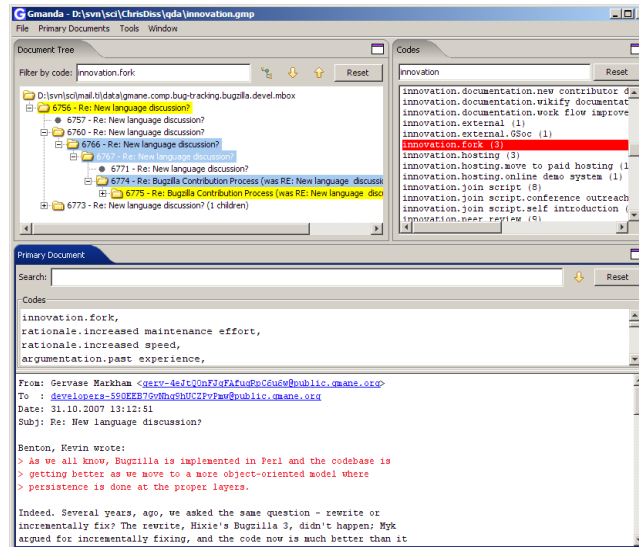


## **Research Ethics for Studying Open Source Projects**

Christopher Oezbek  
Freie Universität Berlin  
23.02.2008

Hello, I am... (see [www.cs.fu-berlin.de/~oezbek/](http://www.cs.fu-berlin.de/~oezbek/) for more information about me).

Today I would like to share with you some thought on Open Source research in general.



I am currently studying mailing list data using a qualitative research approach as part of my research into innovation introduction...

and by qualitative I mean that I read emails...

(The screenshot you see here is of the tool I am using to to aid me with the process of coding each relevant email using Grounded Theory methodology)

Maintainer to committer:

*„...next time I see crap like this I'm going to...  
[...]  
I'm going to go postal on the next  
maintainer who doesn't understand what "merge  
window" and "fixes only" means."*

so for instance, I am looking at the role of the maintainer in the innovation process and so I find emails like this one...

(read message aloud)

and of course this is interesting as it highlights one possibility of what can happen in a team process, when introduced processes get violated by contributors and hints at the „mighty maintainer syndrome“.

And so the question comes up that triggered this paper...

Can I publish this?

Can I publish this?

Because we all know...



The Google logo, consisting of the word 'Google' in its characteristic multi-colored font.

[Advanced Search](#)  
[Preferences](#)**Web**Results **1 - 1** of **1** for "[next time I see crap like this I'm going to](#)". (0.05 seconds)[Re: \[GIT PULL\]\[RETRY\] KVM updates for Linux 2.6.21-rc2 \[LWN.net\]](#)6 Mar 2007 ... I want \*FIXES\*ONLY\* by now, and **next time I see crap like this I'm going to ignore it until you fix me a pull that contains fixes only, ...**[lwn.net/Articles/224886/](http://lwn.net/Articles/224886/) - 13k - [Cached](#) - [Similar pages](#)

Once you type the quote into Google you get exactly one hit, which tells you that Linus Torvalds is the maintainer who wrote this.



From a paper about introducing an information manager role at GNU Classpath:

*„I do not like wikis.“, which is not further elaborated upon by the author.“*

So that got me started thinking... and I went back to previous research and scanned it for similar cases. And in this paper we paraphrased the authors responses, so I tried to see how difficult it would be to find the author without an exact quote...



Google™

classpath "I do not like wikis."

Search

[Advanced Search](#)  
[Preferences](#)

Web

Your search - **classpath "I do not like wikis."** - did not match any documents.

Click to add text

a direct search did not lead to any results... but just 60 seconds of tinkering with the search string found...



The Google logo, consisting of the word 'Google' in its characteristic multi-colored font.

[Advanced Search](#)  
[Preferences](#)

Web

Results 1 - 5 of 5 for classpath "[I hate wikis.](#)". (0.37 seconds)[Re: Get ready for FOSDEM 2006](#)

classpath. <- Thread ->; <- Date -> ... [I hate Wikis](#). OK, I'll do it. :-( Andrew. Get ready for FOSDEM 2006 Mark Wielaard ...

[www.mail-archive.com/classpath@gnu.org/msg11949.html](http://www.mail-archive.com/classpath@gnu.org/msg11949.html) - 6k - [Cached](#) - [Similar pages](#)

Click to add text

the email, that is the origin for the sentence.

interestingly enough, this email was part of the preparations for FOSDEM 2006 in which the participants of GNU Classpath used the wiki that was created as part of the information manager to coordinate their meeting here.

If individuals can be identified easily from research results...

So back to the main question for this talk...

If individuals can be identified easily from research results...

<Next Slide>



## Is there an ethical problem?

Is there an ethical problem to publish these results.

But before I continue with this question, I would like to make this more interesting to the rest of you.

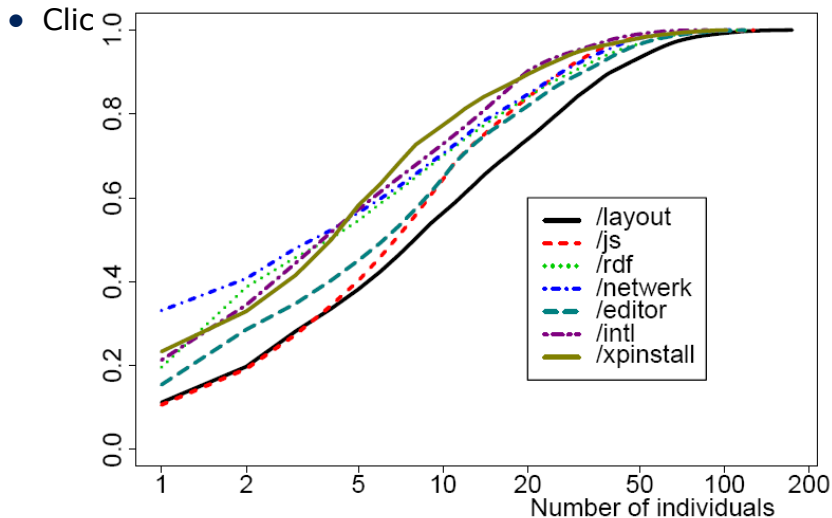
Who of you is doing empirical research with Open Source projects that is not qualitative?



## How about the other research methods in Open Source?

In the paper I have discussed this with a little bit more detail, but here I would like to focus on the two other big two, that I see:

Data mining and survey research.



Mockus et al. 2002: Two Case Studies of Open Source Software Development: Apache and Mozilla

Christopher Özbek, oezbek@inf.fu-berlin.de

12

For data mining I have brought along a plot from the Mockus' Apache vs. Mozilla case study.

This plot shows the cumulative developer contribution to several modules from the Mozilla web browser.

And notice the logarithmic scale for the contribution that makes it easily possible for the first 5 developers to pick out their relative contributions. How hard would it be to find out who developer number 2 is?

**Table 2: Respondent's primary open source projects**

Multiple occur.	Single occurrences			
Linux (21)	AbiWord	Genes	LTSP	Tabindex
Midgard (4)	Analog	Gtk	MPLS for Linux	The COG Engine
Perl (3)	Cons	Hover Carnage	NetBSD	Vaxbb

[Hars and Ou, 2001: Working for Free? – Motivations of Participating in Open Source Projects]

Christopher Özbek, oezbek@inf.fu-berlin.de

13

Surveys have a similar problem. If you do not aggregate data sufficiently it might be possible to pick out somebody. For instance in the above example, some of the respondents came from projects that are so small, that it is likely that the maintainer is the respondent.

So with this I just wanted to show that also the other research methods might allow us to identify participants and thus back to the question, whether it is a problem and in particular a problem of ethical research if it is possible to identify somebody...

So the next step for me, was to look around how other people have been thinking about this problem.

Respect for Persons

Justice

Benefice

So if you look around, you will find the Declaration of Helsinki or the Belmont report and which talks about research with human subjects using 3 general principles:

Autonomy, Benefice and Justice.

Autonomy (Participants must be autonomous during the research, they must be volunteers, must be informed about what is happening with them, must agree to what is happening to them, must be able to cancel their participation)

Benefice:

What can be gained from the research?

And this is an interesting theme that the organizers have chosen for this workshop. How can our research be beneficial?

And the second question, what is the potential damage to participants?

Justice:

Is cost / benefice distributed in a fair manner? Or does one person or one group of people have to carry costs for the general public?

Universities especially in the United States have taken these principles and created a set of rules and review boards around them to make them usable in general day to day science. For instance it is common practice to require an Informed Consent from all participants in the experiment, in which they agree to have understood the cost and benefit of the experiment unless the data is available publicly.



## Discussion Point: What kind of data is public?

So I would like to put these in a series of questions for discussion:

First: What kind of data do we believe to be public enough to be outside the scope of most ethics reviews?



## Discussion Point: What might be damaging to participants?

Second: What do you think is information that we should not publish about participants?

A classic scenario of these damages is certainly the publication of relative performance data. If for instance data mining reveals one project to be the worst with regards to QA, isn't it plausible for the maintainer of this project to have to fear that this information causes him harm on the job market?

What kind of damages do you believe we should protect against? That is which kind of information should we not publish?



## Discussion Point: Ethics review at your university?

So in the beginning I talked to several people in psychology and told them what I think the consequences are and they told me: „But this is ridiculous, we are doing research“ making it sound like censorship.

So I would like to ask you about your experiences. Did you have to go through review, what kind of requirements did you receive?



## Discussion Point: What can we / must we do?

And lastly, let's over-dramatize here a bit: Let's say that not only has Sourceforge been over-fished for surveys participants but let's also say we get a privacy scandal with a research study in the near future which splits research and community.

I do not think this will happen and efforts like this workshop go much to show that we are actually on the other way of a better understanding of research and community. But nevertheless

What can we / should we / must we do? Let's say in a „proactive“ sense.

For instance should we think about an opt-in or opt-out model for survey participation or a register of all research currently conducted similar to the ones used with medical research already?

Thank you!

- What kind of data is public?
- What might be damaging to participants?
- Ethics review at your university?
- What can we / must we do?

Thanks for listening to this talk.

If you have comments on this subject, please do not hesitate to write an email to [oezbek@cs.fu-berlin.de](mailto:oezbek@cs.fu-berlin.de).