# FLOSS 2013: A Survey Dataset about Free Software Contributors: Challenges for Curating, Sharing, and Combining

Gregorio Robles
Universidad Rey Juan Carlos
Madrid, Spain
grex@gsyc.urjc.es

Laura Arjona Reina
Universidad Politécnica de Madrid
Madrid, Spain
laura.arjona@upm.es

Alexander Serebrenik
TU Eindhoven
Eindhoven, The Netherlands
a.serebrenik@tue.nl

Bogdan Vasilescu
TU Eindhoven
Eindhoven, The Netherlands
b.n.vasilescu@tue.nl

Jesús M. González-Barahona
Universidad Rey Juan Carlos
Madrid, Spain
jgb@gsyc.es

## ABSTRACT

In this data paper we describe a data set obtained by means of performing an on-line survey to over 2,000 Free/Libre/Open Source Software (FLOSS) contributors. The survey includes questions related to personal characteristics (gender, age, civil status, nationality, etc.), education and level of English, professional status, dedication to FLOSS projects, reasons and motivations, involvement and goals. We describe as well the possibilities and challenges of using private information from the survey when linked with other, publicly available data sources. In this regard, an example of data sharing will be presented and legal, ethical and technical issues will be discussed.

## Categories and Subject Descriptors

K.4.0 [**Computing Milieux**]: Computers and Society—*General*

## General Terms

Human Factors

## Keywords

Survey; free software; open source; ethics; data sharing; data combining; privacy; anonymization; microdata; open data;

## 1. INTRODUCTION

In 2002, the first FLOSS survey was launched [3]. With over 2,500 participants, it was the first large survey of Free/Libre/Open Source developers around the world. This survey had major impact on the FLOSS community, but also in academia (over 400 citations of it can be found in Google Scholar) and even politics. Because of its success, the survey was replicated in the U.S. [2] and in Asia and Japan [8].

10 years later the survey has been replicated in order to see how the community has changed. The questions in this new survey are almost the same in 2002, although some minor changes have been introduced, as community and circumstances have changed during these years. For instance, the target population included all contributors to FLOSS projects: developers, translators, documenters, community managers, etc.

The goals of this paper are following:

1. To offer a curated data set with data from over 2,000 FLOSS contributors.
2. To present a case study, the challenges and issues of an "augmented" use of the data together with public data from other sources [10].

The relevance of our contribution lies in the fact that data obtained by means of surveys with research purposes is seldom shared. One of the reasons for the lack of sharing is that these data sets contain private data or personally identifiable information. However, much of the information obtained by means of a survey is very difficult (if not impossible) to obtain by other means. Linking data obtained from surveys with other data, gathered by traditional mining software repositories means, may provide new insights and allow for further discoveries.

The paper is structured as follows: in the next section, the survey methodology and the survey questions presented. Then, the dataset is described. Section 4 discusses some limitations of the survey, while Section 5 provides some results that can be obtained from the survey. Next, the possibility of "augmenting" the dataset with other sources is presented, together with the issues and challenges of doing so. Finally, conclusions are drawn.


*MSR* '14, May 31 – June 1, 2014, Hyderabad, India

## 2. SURVEY METHODOLOGY

The survey methodology of the FLOSS 2013 survey has been the same as the one of the original FLOSS survey: an open web-based survey, where participants are self-selected. To attract participants the survey was announced via FLOSS community news sites (such as Slashdot), spread through Twitter and other social networks, and sent to FLOSS mailing lists. The survey opened November 12$^{th}$ 2013 and closed December 6$^{th}$ 2013 and was managed using a libre software called LimeSurvey.

Together with the survey, a website[1] with additional information was offered: the goals, a privacy statement and information on how to contact the research team. The website and the survey questionnaire were available in English and Spanish. Support was provided via Twitter, with a dedicated e-mail address, and with a channel on an IRC network.

The survey could be answered anonymously; IPs were not tracked and no cookies were used. All the questions were optional except the first one, about the type of contribution to FLOSS projects, as it was a branching question. Participants were asked to provide their e-mail address (or some part of it) in order to ascertain that they really were FLOSS contributors. They were informed that while the rest of non-private information would be available publicly, e-mail addresses would be handled as private data and will not be made public nor shared with other research groups.

The 58 questions can be classified into following areas[2]:

- Personal situation (gender, civil status, number of children, country of birth and of residence/work)
- Education (highest level of education, level of English)
- Professional situation (profession, satisfaction, income)
- FLOSS perspective (free software vs open source)
- Development (age when joining FLOSS, reasons and motivations for joining, reasons and motivations for still participating, earn money with FLOSS)
- Technology (favorite editor, programming languages)
- Economic and community rewards (job opportunities, expectations from other developers, challenges)

## 3. DESCRIPTION OF THE DATA

An anonymized version of the survey results is available publicly, under a Creative Commons CC-By license, in the survey's website[3]. The data is downloadable by now in two formats: one for R (a schema for importing the data into R and its corresponding CSV datafile) and another one for importing the data into LimeSurvey.

The schema of the data is very simple as it is composed of a single table, with one row per response. The total number of columns is 263, as many questions have multiple responses.

As the questionnaire was split in several web pages, we obtained a higher number of responses for the questions that appeared first. Table 3 shows how the number of responses decreases for some selected questions. The dataset includes 1,644 complete submissions (where the participant viewed all the questions), plus 539 incomplete submissions (those participants left before reaching the last page), making a total of 2,183 records.

**Table 1: Number of answers for some of the questions of the FLOSS 2013 Survey.**

| Question | # of Responses |
|---|---|
| Type of contribution | 2,183 |
| Gender | 2,035 |
| Income | 1,818 |
| Number of projects | 1,795 |
| Complete questionnaire | 1,644 |
| Total responses | 2,183 |

Inspecting the emails provided, we found several cases of duplicated responses (i.e., two questionnaires with the same email). In those cases we have removed the duplicates, keeping the questionnaire that has a maximum number of answers. In total, 50 responses out of 1,526 responses with complete email address have been discarded. In the rest of questionnaires a similar duplication rate (of around 3%) should be considered.

## 4. THREATS TO VALIDITY

The most important threat to the validity of the data is that a non-random sample population has been used, resulting in a biased sample. For instance, we know that the survey has been promoted strongly among outreach programs for women in open source, which know well the results of the FLOSS 2002 survey regarding to gender topics. This may have had an impact in the number of women actively promoting and participating in the FLOSS 2013 survey. Bethlehem argues that when performing a self-selected web survey, "[t]he underlying question is whether such a survey can be used as a data collection instrument for making valid inference about a target population. [...] This seems to be similar to the effect of nonresponse in traditional probability sampling based surveys" [1]. In the case of this survey, the selection bias can be specially skewed for demographic data. However, statistical techniques, such as the Heckman correction [7], can be used to correct the bias, especially in cases such as this survey where the sample is very large.

Nevertheless, the survey is representative in many aspects, in particular those that are not related to demographic information. It also may serve to correlate demographic data with other type of data, such as to study if there are gender differences in the type of contributions or in the type of motivations for participating in FLOSS projects.

## 5. SOME RESULTS

We want to remark three of the possibilities that the FLOSS survey dataset offers: (1) It is possible to update the statistics and reports generated by the FLOSS Survey 2002. A comparison between the two surveys may provide a notion of how the FLOSS community has changed in the last decade. (2) Since the dataset provides information about non-coding contributors, it is possible to perform comparisons between

[1] `http://floss2013.libresoft.es`

[2] The complete questionnaire, including answers, can be obtained from `http://floss2013.libresoft.es/downloads/questions/FLOSSSurvey2013_en.pdf`

[3] `http://floss2013.libresoft.es/results.en.html`

them with FLOSS developers. (3) A wider sample of female contributors produces more reliable assessments about the participation of women in open source.

We will briefly show an example of a study that could be done with the data. Given that later we will use gender data, the examples will consider that aspect in particular. So, for instance, Figure 1 gives, for males and females, the age contributors had when they started collaborating in FLOSS projects. It can be observed how females join FLOSS at older ages.

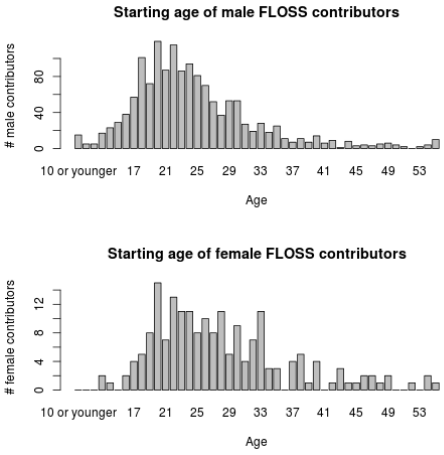


**Figure 1: Age respondents had when they first contributed to FLOSS (grouped by gender).**

On the other hand, Figure 2 shows graphically the year contributors started to collaborate, again grouped by gender. While male respondents seem to have stabilized since the early 2000s, the number of female FLOSS contributors presents an increase in the last 5 years.

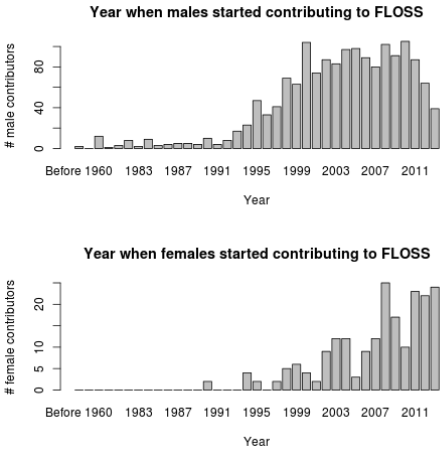


**Figure 2: Year the first contribution to FLOSS took place (grouped by gender).**

# 6. COMBINING DATA, AND PRIVACY

In the field of mining software repositories, the amount of public data is vast. However, this data is usually related to the development artifacts (files, packages) or process (logs, emails, traces). Data such as the one that is offered by the FLOSS surveys is difficult to obtain from public sources.

Combining data sources may provide researchers with an *augmented* view of the matter of study. Sometimes some demographic variables affect the results of an investigation, as it is well known from other fields of research; so, for instance, in the educational research with online games significant differences have been identified depending on the gender of the students [12].

However, sharing and combining data supposes several risks, and traditional anonymization techniques have proven to be limited [6]. Legal, ethical and practical issues have to be considered. For example, in the European Union, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data regulates the processing of personal data regardless of whether such processing is automated or not[4].

Offering the results in an anonymized and aggregated way may help on this. But even this has to be done with care as survey respondents may be recognized from the output[5]. For example, we could anonymize identities by using a sequential id, known only to us, for each respondent. But if we publish this anonymized data with detailed commit counts, identities may be inferred for a certain number of cases, because the commit count can be easily tracked for certain projects.

There are several approaches that have been proposed to ensure *secure* anonymization and that we would like to study in the next future. The concept of k-anonymity [9] tries to ensure that with k-anonymity greater than 1, even with all fields a single person cannot be identified, but k people. Still, k-anonymity has shown not to be sufficient as attackers can discover sensitive attributes in data with low diversity, and together with other information identify a single person. Data with sufficient diversity, l-diversity, should be published [5]. Finally, t-closeness requires that the distribution of sensitive attributes to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should less than a threshold t) [4]. In the meantime, we will combine the data internally, as we have done in the case study shown next.

# 7. CASE STUDY OF COMBINING DATA

Our aim is to link the FLOSS survey data with data from other sources, in this case data from StackOverflow, to show its potential uses. StackOverflow is the largest Q&A website for programmers, with more than 2.3M users registered as of September 2013. New StackOverflow data dumps are available quarterly in XML format.

To merge the two datasets of survey respondents and StackOverflow users, we followed a conservative approach and made use of email addresses. In the FLOSS dataset email addresses are present, while for StackOverflow users only anonymized versions of the email addresses (MD5 hashes) are available. Therefore, we linked a FLOSS respondent

[4]Currently, a draft of a new regulation that will supersede this directive is under discussion by the European Commission.

[5]See *Ars Technica*: "Anonymized data really isn't - and here's why not", `http://goo.gl/ibuVp`.

**Table 2: Measures when combining StackOverflow gender resolution results with the FLOSS 2013 survey.**

| Gender | Male | Female | Total |
|---|---|---|---|
| Precision | 0.97 | 0.55 | 0.90 |
| Recall | 0.54 | 0.39 | 0.52 |
| F-measure | 0.69 | 0.46 | 0.66 |
| MCC | 0.26 | 0.42 | 0.62 |

to a StackOverflow user if the computed MD5 hash of the former's email address was identical to the MD5 email hash of the latter. Similar approaches have been used successfully in the past [11].

To automatically infer gender for StackOverflow users, we used a previously-validated [10] name-based gender resolution tool. The tool[6] tries to infer a person's gender based on their name and, if available, their location. It uses lookup tables with first names for different countries (e.g., Andrea is a common male first name in Italy, but a common female one in Germany) as well as a number of heuristics (related, e.g., to gender-specific last name forms, cross-country lookup, or diminutive resolution). The samples are composed of 1,476 FLOSS survey respondents that provide a complete and valid (at least from its construction) e-mail address, which has been hashed with MD5. For StackOverflow, we have 2,091,063 distinct MD5 hashes of e-mail addresses out of a total 2,332,406 total MD5 hashes gathered.s As a result of matching the MD5 hashes in both datasets, we have obtained 451 matches. From these, 439 had provided gender information in the FLOSS survey. Considering the gender resolution algorithm used with StackOverflow, we have identified 227 correct gender matches.

Table 2 provides further overview of the result of our validation. Traditional information retrieval measures, such as precision, recall and F-measure have been obtained. In addition, we provide the Matthews correlation coefficient (MCC). This coefficient is a measure of the quality of a binary classification and is seen as a balanced measure which can be used even if the classes are of very different sizes. As binary classification we use male/non-male, female/non-female and success/failure (for the "Total" column) in the MCC row in Table 2. Values of MCC are between -1 and +1, representing +1 a perfect prediction and 0 no better than a random prediction. The formula of the MCC is:

$$MCC = \frac{t_p * t_n - f_p * f_n}{\sqrt{(t_p+f_p)(t_p+f_n)(t_n+f_p)(t_n+f_n)}}$$

where $t_p$ stands for *true positives*, $t_n$ for *true negatives*, $f_p$ for *false positives* and $f_n$ for *false negatives*.

# 8. CONCLUSIONS

In this data track paper, we offer data from an open, self-selected web survey on FLOSS contributors that has been answered by over 2,000 participants. We have presented the methodology that has been followed to gather the data, briefly described the data and showed the threats to validity of the survey. Some of the results that can be obtained from the data have also been shown.

[6] `https://github.com/tue-mdse/genderComputer`

Then, we have presented one of the possibilities that such data provides: combining it with other public data obtained by means of mining software repositories (MSR). We see that our data is specially suited to *augment* these studies, as it offers valuable data that is difficult to obtain by traditional MSR means. We discuss some possibilities to still perform such types of analysis by means of a case study. However, we show that because of having private data, in our case e-mail addresses, this is a task that has legal and ethical challenges that still need to be considered in future research.

# 9. ACKNOWLEDGMENTS

The work of G. Robles and J. M. González-Barahona has been funded in part by the Spanish Government under project SobreSale (TIN2011-28110). The research of B. Vasilescu has been financed by the Dutch Science Foundation with project NWO 600.065.120.10N235.

# 10. REFERENCES

[1] J. Bethlehem. How accurate are self-selection web surveys? Technical Report Discussion paper (08014), Statistics Netherlands, 2008.

[2] P. A. David, A. Waterman, and S. Arora. FLOSS-US: The Free/Libre/Open Source Software Survey for 2003. Technical report, Stanford Institute for Economic and Policy Research, Stanford, USA, 2003.

[3] R. A. Ghosh, G. Robles, and R. Glott. Software source code survey (free/libre and open source software: Survey and study). Technical report, Univ. of Maastricht, The Netherlands, June 2002. http://www.flossproject.org/report.

[4] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, volume 7, pages 106–115, 2007.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[6] P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6), 2010.

[7] P. Puhani. The Heckman correction for sample selection and its critique. *Journal of economic surveys*, 14(1):53–68, 2000.

[8] H. Shimizu, J. Iio, and K. Hiyane. The realities of Free/Libre/Open Source Software developers in Japan and Asia. *First Monday*, 9(11), 2004.

[9] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[10] B. Vasilescu, A. Capiluppi, and A. Serebrenik. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, page iwt047, 2013.

[11] B. Vasilescu, V. Filkov, and A. Serebrenik. StackOverflow and GitHub: Associations between software development and crowdsourced knowledge. In *Proceedings 2013 Intl Conf on Social Computing*, pages 188–195. IEEE, 2013.

[12] H.-Y. Wang and Y.-S. Wang. Gender differences in the perception and acceptance of online games. *British Journal of Educational Tech*, 39(5):787–806, 2008.