

Free/Libre and Open Source Software: Survey and Study

FLOSS

Deliverable D18: FINAL REPORT

Part 4A: Survey of Developers – Annexure on validation and methodology

Rishab A. Ghosh

Ruediger Glott

Bernhard Krieger

Gregorio Robles

International Institute of Infonomics
University of Maastricht, The Netherlands

October 2002

A note on developer survey methodology

(In response to comments received after the initial publication of the survey report)

Early in the design of the survey methodology we faced the question of sampling: was it possible to survey a representative sample of developers? The question actually has two parts: is it possible to ensure that respondents are developers, and is it possible to identify a sample that is representative of developers based on some filtering criteria?

Our conclusion was that there is insufficient empirical data on OS/FS developers to identify the criteria of sampling. However, without empirical data as a basis it is not possible to demonstrate that a chosen sample of respondents is representative of developers in general: i.e. it isn't possible to sample developers and know with any confidence that the distribution of nationalities, skill levels, income levels or leadership is representative of the distribution in the total (unsampled) population of developers.

Therefore, we decided that in order to have results that would be empirically valid, we would have to have a random sample. The survey was self-distributing – i.e. it was posted to various developer forums, and then re-posted by developers to other forums, some of which are listed on the workshop website. The survey announcement was translated into various languages, to correct possible biases inherent in an English-language survey. We can state that the survey was seen by a very large number of developers (it was announced on slashdot, among other places) and therefore the sample that chose to respond was random, though with some identifiable bias. We are also able to state that the respondents are developers, through the validation process described below.

There is a self-selecting bias that is inherent in any survey that is answered voluntarily. There is also a possible bias towards more motivated developers, though not necessarily the more *politically* motivated. There may be some over-representation of some nationalities – e.g. France has not previously appeared so high on demographic surveys of OS/FS developers. One reason for this could be that the survey announcement was translated into French (and distributed on French developer forums). Similarly, there may be under-representation of some nationalities especially Asian developers.

Validation of survey responses

One of the problems aligned with online questionnaires is that often it cannot be verified whether the respondents really belong to the group that is under scrutiny. Therefore, to validate our data in this respect, we used a sub-sample of 487 respondents individually identified as OS/FS developers as a measure to validate our overall results by comparing their answers to the answers of those who were not known personally. (Developers were identified by matching the “e-mail address” field to names found in the source code analysis – see part 5 of the final report – or matching them to sources on Internet archives.)

The validation simply consisted of a comparison of means and standard deviations of the two groups (“known developers” and other respondents) with regard to a selection of variables of our data set. The comparison is presented in the tables at the end of this annexure, which highlight the few relatively large differences between verified and non-verified responses. This shows that the group of verified OS/FS developers consists of slightly more active and “professionally” experienced persons, but their answers do not differ significantly from those of the non-verified OS/FS developers, especially in terms of orientations and motivations.

The whole procedure of the validation kept, of course, to the privacy requirements of the respondents. The first step, identification of the sub-sample, was conducted separately from the main analysis of the survey data. Only the ID-number (a serial number uniquely generated for each respondent) and two variables providing personal features (e-mail address / address fragment) were used for the first step. All other data based on the answers given to the FLOSS online questionnaire was excluded from this process. After identification, the data of the sub-sample were made anonymous by replacing all personal information by the single attribute “verified” or “not verified”. After this transformation, the validation data were integrated into the data set of the survey containing information about the answers given to the questions of the online questionnaire. Thus, at no point of the analysis was it possible to assign answers to the questions of the survey to particular persons as identified through source code or online archives.

Comparative table of verified and non-verified respondents

Variable	Not verified developers	Verified developers	Not verified developers	Verified developers
	Mean		Std. Deviation	
OSORFS Part of FS or of OS community	1,7	1,7	0,8	0,8
OSFSDIFF Differences between OS and FS community	1,9	1,9	0,7	0,7
STRTYEAR Start year FLOSS development	47,8	47,3	4,0	4,5
STRTAGE Start age FLOSS development	14,0	13,3	6,2	6,3
HRSOSFS Hours per week spent in FLOSS development	2,7	3,4	1,5	1,5
ALSOPROP Also developing proprietary software	1,5	1,5	0,5	0,5
YEARPROP Start year developing proprietary software	20,7	23,2	23,3	23,2
HRSROP Hours per week spent in developing proprietary software	4,2	4,1	1,6	1,6
USEOSFS Usage of OS/FS at work, university or school	1,1	1,0	0,2	0,2
MAININC Main income earned by development, support, or administration	2,8	2,8	1,0	1,0
KINDOSFS Primarily developed kind of OS/FS	2,9	3,5	1,6	2,0
AREAOSFS Area for which OS/FS is primarily developed	5,2	5,0	2,2	2,2
DISTRSYS Favorite distribution/operating system	6,8	6,3	5,3	5,5
DESKTOP Favorite desktop	2,9	2,7	2,0	2,0
EDITOR Favorite editor	1,9	1,8	0,8	0,8
PROJALL Number of all OS/FS projects so far	1,4	1,7	1,1	0,8
PROJCURR Number of current OS/FS projects	3,0	3,9	1,6	1,4
LEADER Number of OS/FS projects involved as leader, administrator, or coordinator	2,1	2,7	1,4	1,3
CONTACTS Number of regular contacts to other OS/FS developers	3,0	4,0	2,5	2,0
SATISFY OS/FS satisfying todays requirements for software better than proprietary software?	1,6	1,6	1,1	1,0
TIMEPRES Developing Software is usually associated with time pressure	2,5	2,5	1,0	0,9
JOYFUL Working in this field is joyful	1,7	1,5	1,1	1,3
BORING Working in this area can be very boring	2,9	3,0	1,0	1,0
EFFICIENT The organization of work in this area is much more efficient	2,5	2,5	1,5	1,5
QUALITY The developed software is of high quality	2,1	2,2	1,5	1,4
AESTHET People write beautiful and aesthetic programs	2,1	2,1	1,5	1,4
INNOVAT Innovations are made in this area	2,7	2,9	1,4	1,5
COMMONOS People in OS/FS are more concerned about money than in proprietary sw domain	2,9	2,9	0,4	0,5
COMMONPR People in domain of proprietary sw domain are more concerned about money than in OS/FS domain	1,9	1,9	0,7	0,7
EXPERTIS Positive impact of experience in OS/FS on job opportunities?	1,0	1,2	0,6	0,8

(continued)

Variable	Not verified	Verified	Not verified	Verified
	developers	developers	developers	developers
	Mean	Mean	Std. Deviation	Std. Deviation
BALANCE1 Assessment of own balance in OS/FS scene	2,5	2,5	1,1	1,1
BALANCE2 Assessment of others' balance in OS/Fs scene	3,1	3,1	1,7	1,8
SELFEXPL OS/FS as a kind of self exploitation	2,6	3,5	0,8	1,7
MARKCODE Marking sourcecode as yours?	1,5	1,4	0,6	0,6
GENDER What is your Gender	1,0	1,0	0,1	0,1
PARTNER Do you have a partner?	2,2	2,2	1,2	1,2
CHILDREN Do you have children?	1,2	1,2	0,4	0,4
CHILNUMB Number of children	0,2	0,3	0,9	0,7
NATIONAL Nationality	10,7	11,5	6,3	6,5
RESIDENC Country of Residence or work	10,8	11,7	6,2	6,4
PROFESS Profession	6,9	6,5	5,9	6,0
JOBSATIS Job satisfaction	1,4	1,8	1,1	1,2
EDUCAT Highest level of education	4,7	4,6	1,7	1,6
MOTONGUE Mother tongue	3,8	4,9	4,2	4,4
INCOME Monthly gross income	3,7	4,0	2,0	1,9
STATUS Employment status	2,3	2,3	1,1	1,1
PROF_1 Professions (general)	3,4	3,2	2,2	2,3
PROF_2 Profession in IT-Sector or Other Sectors	1,2	1,1	0,4	0,4
AGE	23,0	22,3	6,2	6,3
AGENOW Current age	27,1	27,1	7,1	7,2
DURATION Duration of participation in OS/FS scene	4,2	4,8	4,1	4,3
STYAR_R Startyear FLOSS development - grouped	5,7	5,2	2,2	2,4
YARPRO_R Startyear proprietary software development - grouped	2,3	2,5	3,2	3,1
STRAGE_R Startage FLOSS development - grouped	3,6	3,4	1,4	1,4
MIGRAT Migration of OS/FS developers	1,1	1,1	0,3	0,3
NATION_R Nationality - grouped	1,5	1,5	0,8	0,8
RESID_R Residence - grouped	1,4	1,5	0,8	0,7
PROJAL_R Number of all OS/FS projects so far	1,3	1,6	0,7	0,6
PROJCU_R Number of current OS/FS projects	2,9	3,7	1,2	1,2
LEAD_R Leadership experience in OS/FS projects	2,1	2,6	1,2	1,2
CONTAC_R Number of regular contacts to other OS/Fs developers	2,8	3,5	1,5	1,4
YEAR Startyear OS/FS development	1996,8	1996,3	4,0	4,5
PROPSTRT Startyear proprietary software development	899,0	1013,7	998,2	992,8
DURAT_R Duration of membership in OS/FS community	3,1	3,6	2,0	2,1
STATUS_R Employment status - grouped	1,2	1,2	0,4	0,4
AGENOW_R Current age - grouped	3,3	3,3	1,6	1,5