Women in Debian 2013

Laura Arjona Reina, Patty Langasek, Miriam Ruiz

August 2013

1 Abstract

In June, 2005, Magni Onsøien performed a study about women in Debian and free software, analyzing the participation of people in several mailing lists and providing aggregated data grouped by gender. Gender of participants was guessed by manual inspection of the email sender's names. She compared her results with the FLOSS Survey of free software developers performed in 2002. The main conclusions were that "both debian-level and the Debian Developers have about only half the expected amount of women if compared to these studies and other comparable projects". On the other side, she found that there were many women participating in debian-women sub-project, showing that "there are in deed women interested in Debian, and it is possible to find them."

Since then, several actions promoting diversity in free software projects in general, and in Debian in particular, have been taken in the free software scene. In addition to this, many free software projects have gained popularity (being one of them the Debian project, and the Ubuntu operating system, a Debian derivative) along with the open source development model becoming "mainstream". Free software developers and communities have begun to use mainstream communication systems as social networks, and the tools to develop free software have been improved and simplified, lowering the access barrier to become part of the development team.

We can suppose that after all this, the participation of women in Debian has increased, and the gender imbalance is not so skewed now. This study pretends to repeat Onsøien study, in order to confirm (or not) this supposition.

2 Old study's data

Magni Onsøien analyzed few developers' mailing lists during three months.

The gender classification of senders was male, female, unknown (uncertainty about the gender of a person), and uncathegorized (not enough data to decided gender, e.g. only email address, nick, or initials).

The analyzed lists were the following: debian-devel, debian-women, debian-vote, httpd-dev, gentoo-devel, gnome-desktop, gnome-devel, kde-core-devel, kde-devel, linux-kernel.

The period analyzed was November 2004 - January 2005 (March - May 2005 for the Linux kernel lists).

In addition to this, statistics about total number of developers and number of female developers were given for Debian, the GNOME Foundation and Gentoo, and number of women in the Debian New Maintainer queue.

3 Old study's results

3.1 Numbers (summary)

In table 1 we can find the participation of women in Debian mailing lists in 2005.

List	Number of senders	Women %	Men %
Debian-devel	844 senders	0,83% women	73,82% men
Debian-vote	129 senders	2,32% women	90,69% men
Debian-women	103 senders	40,78% women	51,46% men

Table 1: Women in Debian 2005 (Magni Onsøien's results)

3.2 Conclusions

The main conclusions were that "both debian-level and the Debian Developers have about only half the expected amount of women if compared to these studies and other comparable projects". On the other side, there were many women participating in debian-women sub-project, showing that "there are in deed women interested in Debian, and it is possible to find them."

4 New study: Data sources and methodology

We are going to analyze the Debian mailing lists in a similar period of time, and compare results with 2005.

We'll try to replicate the same study but with a new free software tool: MLStats from MetricsGrimoire suite.The MetricsGrimoire suite[1] is a complete set of tools for analyzing free software communities, developed by the LibreSoft research group, the Bitergia company, and the MetricsGrimoire community. MailingListStats (MLStats) is a free software tool written in Python that parses a mbox file into a MySQL database, storing the information of senders, mailing list and messages in different tables. On the resulting "people" table we will perform the gender classification by manual inspection and internet search, with the help of the database of senders that Magni Onsøien created, and two automatic methods for gender classification by name: the program gender.c and the US Census name database.

As a side conclusion of this study, we will find out if the automatic means for gender classification are accurate for this kind of study or not.

[1] http://metricsgrimoire.github.io/

4.1 Mailing lists sender retrieval, parsing into a SQL database

These are the lists and periods that Magni studied and the periods that we are going to parse for the new study:

List	Magni Onsøien's studied period	New study's period
Debian-devel	Nov 2004 - Jan 2005	May 2012 - Nov 2012
Debian-vote	Feb and Mar 2005	Feb and Mar 2013
Debian-women	Mar - May 2005	Jun 2012 - Jun 2013

Table 2: Studied mailing lists periods

Debian does not provide mbox files with its mailing list archives (although Debian Developers have access to them). However, the Gmane service which also stores the mailing lists, provides an "export" utility which allows to download a mbox file given the initial and final mail of the list.

- Gmane provides mbox format downloads for the debian mailing lists, but you have to provide the initial mail and the last mail that you want to be saved.
- We decided to study from June 2012 to June 2013 for debian-women and debian-devel lists (the debian-women mailing list has become a low-traffic list in these years, so we have chosen the whole year as period for the study), and February - April 2013 term for debian-vote (around DPL elections).
- We determined the initial and last email to be downloaded by manual inspection.
- We download the mboxes with wget.

```
wget http://download.gmane.org/gmane.linux.debian.women/4468/4623+
wget http://download.gmane.org/gmane.linux.debian.devel.vote/16282/16674+
wget http://download.gmane.org/gmane.linux.debian.devel.general/
            173451/184670+
```

We create a folder "mailboxes" and 3 subfolders, one for each mail.

We launch mlstats to explore the mailboxes folder, so it will add all the data to the database `mlstats_dw2013`, but each list will be identified:

```
mlstats -db-user larjona -db-password larjona
   --db-name mlstats_dw2013
   --db-admin-user larjona --db-admin-password larjona
   ./mailboxes
```

We clean a bit the table people, removing senders that are bots or official lists addresses:

```
delete from people where email_address like '\%@bugs.debian.org';
delete from people where email_address like '\%@lists.debian.org';
delete from people where email_address like '\%@lists.alioth.debian.org';
```

4.2 Creation of tables about "gender"

We have a text file with the command to create the tables that will store the information about:

```
mysql -u larjona -p < create_tables_gender.sql
```

We fulfill the table gender with the data from table people, splitting names in firstname and familyname:

```
insert into gender (mail, fullname, firstname, famname)
   select email_address, name ,
   substring_index(name, ' ', 1) as firstname,
   substring(name, locate(' ', name)) as familyname from people
   order by name;
```

4.3 Gender classification

4.3.1 Program gender.c

The program "gender.c"[2] is a program for determining the gender of a given fist name. It uses a dictionary file containing a list of more than 40,000 first names and gender, covering the vast majority of first names in all European countries and in some overseas countries (e.g. China, India, Japan, U.S.A.).

We dump the contents of the firstname column into a txt file for checking gender with gender.c

```
select distinct firstname from gender
   order by firstname into outfile '/tmp/firstnames.txt';
```

I have compiled gender.c and generated an executable "gender"

```
./gender -get_gender_for_file ../firstnames.txt ../firstnames_gend
```

The resulting file is in the format

```
name:   XXX
```

[2] http://www.autohotkey.com/board/topic/20260-gender-verification-by-forename-cmd- line-tool-db/

where name is the firstname, and XXX can be ‘is male’, ‘is female’, ‘is mostly male’, ‘is mostly female’, ‘is unisex name’ or ‘name not found’.

I have reformatted the resulting file (using search & replace) to put it in the format

```
name X
```

where X is M, F or U, and name and X are separated by tabs.

Then I have imported the file into mysql

```
load data infile '/tmp/firstnames_gender.txt' into table genderC character
set 'utf8';
```

And then, I fulfilled field “genderC” in table “gender” with that information:

```
update gender, genderC set gender.genderC = genderC.gender where
gender.firstname = genderC.firstname;
```

4.3.2 Magni Onsøien previous work (manual inspection)

Magni Onsøien provided a set of txt files with the following format (separated by spaces)

```
G NN From:  full name <email>
```

Being X the gender (M for male, F for female, - for Unknown, blank for not checked), and NN the number of posts.

I have merged all the files in a txt file and used search & replace in a text editor to provide the following format (separated by tabs)

```
G NN full name email
```

I have removed leading or trailing spaces, and ” characters in fullname strings.

Then I have imported the file in a MySQL table:

```
load data infile '/tmp/total_lists.txt' into table genderM character set
'utf8';
```

I have dropped all the registers without gender:

```
delete from genderM where gender like '';
```

And transformed the '-' gender into 'U':

```
update table set gender = 'U' where gender = '-';
```

Then again, I fulfilled field "genderM" in table "gender" with that information:

```
update gender, genderM set gender.genderM = genderM.gender where
gender.fullname = genderM.fullname;
```

4.3.3 US Census name database

The US Census Government Office[3] provides two files with female names and male names, along with their frequency of appearance in the census.

These files are separated by spaces, I have transformed in tab-separated txt files and imported them into MySQL tables:

```
load data infile '/tmp/dist.male.first.txt' into table genderUM
character set 'utf8';
load data infile '/tmp/dist.female.first.txt' into table genderUF
character set 'utf8';
```

And then, I fulfilled the `genderU` field with that information:

```
update gender set genderU = 'M' where upper(firstname) in (select
firstname from genderUM where firstname not in (select firstname from
genderUF));

update gender set genderU = 'F' where upper(firstname) in (select
firstname from genderUF where firstname not in (select firstname from
genderUM));
```

4.3.4 Determine the gender of each sender by manual inspection.

We dump the contents of table gender in a txt file, for manual inspection of gender:

```
select * from gender order by fullname
    into outfile '/tmp/senders.txt';
```

After that, we need to review by manual inspection all the fields in the table "gender" in order to determine gender, which will be stored in "genderH" field.

[3] http://www.census.gov/genealogy/www/data/1990surnames/names_files.html

Gender classification (”Male”, ”Female”) has been made based on information about name, photo appearance, self-definition in personal webpages (for example, Github personal page, or Twitter profile), or how other people in Debian address that person (for example when advocating him or her as Debian Maintainer). For people in the Debian people database, we will look at his/her registry searching the Debian LDAP database and we will store the gender value (if fulfilled) in the genderH field.

People who self-identify different than man or woman (for example in their personal page) are catalogued as ”Female” for this study. Senders who don't identify their gender and we didn't get a clear information by their name or photo appearance, have been classified as ”Unknown.

Senders corresponding to institutional accounts or spammers have been discarded.

About the internet search for guessing the gender, we search by the fullname plus “debian” or “linux” or “free software” keywords, or by the email field, in order to try to find the personal website of that person, and try to find a photo or a description of him/herself. If we don't find any photo or gender information, we will search the Debian mailing list trying to find emails on that person's application for Debian Maintainer (other people writing about ‘him’ or ‘her’ so we can determine the gender). If we cannot find any photo or self-definition, we guess the gender by the firstame using the automated gender classification tools, or classify as “unknown”.

We will store the date of that research in the “reviewed” field, and from which page or email we obtained the gender in the “notes” field.

After all this process, we can give statistics based on the genderH field.

4.4 Final list cleaning and duplicates removal

We have performed a second clean on the list of senders removing spam and other official accounts.

We need to inspect the list of senders of each mailing list, in order to remove duplicates (senders with same fullname but different email, for example).

We take the Devel list as an example, we have followed a similar process in each list.

- We obtain the senders list. Total: 676

```
select fullname, mail from gender where mail in
    (select email_address from mailing_lists_people
         where mailing_list_url like '\%devel\%')
    order by fullname;
```

By manual inspection we count the duplicates.

Found duplicates: 66 (2 U, 64 male) (See list-duplicates-devel)

Then we obtain the statistics about gender and correct the number subtracting the duplicates:

Raw numbers:

```
select genderH, count(*) from gender where mail in
   (select email_address from mailing_lists_people
     where mailing_list_url like '\%devel\%')
group by genderH;

+---------+----------+
| genderH | count(*) |
+---------+----------+
| F       |        5 |
| M       |      613 |
| U       |       58 |
+---------+----------+
```

Corrected stats for devel list

```
+---------+----------+
| genderH | count(*) |
+---------+----------+
| F       |        5 |
| M       |      549 |
| U       |       56 |
+---------+----------+
Total: 610
```

5 Results

5.1 Mailing lists analysis: stats based on genderH field

Stats for debian-devel list and comparison with Magni Onsøien's results

```
+---------+----------+
| genderH | count(*) |
+---------+----------+
| F       |        5 |  0,82%  (Magni:  0,83%)
| M       |      549 | 90,00%  (Magni: 73,82%)
| U       |       56 |  9,18%
+---------+----------+
Total: 610 (Magni: 844)
```

Stats for debian-vote list and comparison with Magni Onsøien's results

```
+---------+----------+
| genderH | count(*) |
+---------+----------+
| F       |        4 |  6,78%  (Magni:  2,32%)
| M       |       54 | 91,53%  (Magni: 90,69%)
| U       |        1 |  1,69%
+---------+----------+
Total: 59 (Magni: 129)
```

Stats for debian-women list and comparison with Magni Onsøien's results

```
+---------+----------+
| genderH | count(*) |
+---------+----------+
| F       |       25 | 54,35%  (Magni: 40,78%)
| M       |       16 | 34,78%  (Magni: 51,46%)
| U       |        5 | 10,87%
+---------+----------+
Total: 46 (Magni: 103)
```

5.2 Other stats about women participation in the Debian community

- Number of women developers: in 2005, there were 3 females out of 965 DD, 0,31%. In 2013, 18 women out of 988 DD (1,82%), 5 women voted for DPL (27,78% participation), 385 men voted for DPL (39,70% participation).
- One women candidate for DPL (Margarita Manterola, in 2010).
- Number of women in the NM queue: 1 (in 2005, 9 women).
- The Debian wiki hosts some updated statistics [4]: First uploads (58 women), DD account (18 women), DM key (3 women)

[4] https://wiki.debian.org/DebianWomen/Projects/Statistics

6 Conclusions

Women visibility and participation in the Debian community has increased in the last years. Although the participation in the Debian developer mailing list is similar to the participation found in 2005, the numbers about women maintaining packages, women that are Debian developer or that participate in DPL elections are quite higher than before.

As we stated before, the Debian-women mailing list has become kind of low-traffic list. The participation of women there has increased but participation of men in that list is still significant, so we cannot say that the Debian-women list has become "only-women in Debian". I consider this a good sign of diversity and integration in the Debian community.

We cannot trust the automated gender classification methods, human review is still needed, specially if we consider that free software is a global community and participation of people from Asian, South American and Middle-East countries has increased in the last years.

Appendix A: Comparison between manual and automated gender classification

I find out that the percentage of errors is high (specially for women) if we use only the "gender.c" program and/or the US census database to guess people's names. The US Census database is oriented to names used in North America. However, the Debian community is a global community and we find names from very different countries. On the other side, the "gender.c" name database has 42.000 names classified by country, but we have no information about the country in our senders list.

Below we show some number comparing the results of gender classification using the "gender.c" database and the US name database, comparing them with our manual classification (the "genderH" field).

Classification	**genderC**	**genderU**	**genderM**	**genderH**
NULL	28	541	530	0
F	33	28	5	32
M	550	164	195	638
U	122	0	3	63

Table 3: Gender classification with different systems

Error in classification	genderC	%	genderU	%
M AND genderH not M	14	2,55	14	8,54
F AND genderH not F	2	6,06	11	39,29

Table 4: Errors in Gender classification systems

Our main conclusion is that the automated classification can used as one source to be taken into account altogether with other sources. When we didn't find information about that person, we took into account this automated classification to "guess from the name".

Appendix B: Actions taken in the Debian community to increase diversity

Since 2004 with the creation of the Debian-Women team and mailing list, several other actions have taken in the Debian community in order to promote women participation and increase diversity in general. Among them, we can find the following:

- Many of the Debian-women related website information has moved to the main website (www.debian.org/women). There are also pages in the Debian wiki completing that information (https://wiki.debian.org/DebianWomen).
- Debian diversity statement (http://www.debian.org/intro/diversity), written, discussed an approved in 2012.
- The mail alias antiharassment@debian.org has been setup as contact point for Debian event organizers willing to have an anti-harassment policy
- The Debian project has participated in the GNOME Outreach Program for Women (OPW) in 2013 (https://wiki.debian.org/OutreachProgramForWomen). A small report of the first results of this participation is in https://lists.debian.org/debian-women/2013/06/msg00000.html
- Discussion in the 2013 DPL elections about the getting new people into the Debian project (https://lists.debian.org/debian-vote/2013/03/msg00008.html and about the participation of women in Debian (https://lists.debian.org/debian-vote/2013/03/msg00208.html).

- Enrico Zini and Francesca Ciceri have proposed several ideas to increase visibility of the non-packaging Debian contributors (http://www.enricozini.org/2012/debian/more-diversity-in-skills/).

References

- The Debian project - http://www.debian.org
- Debian-Women - http://www.debian.org/women
- Some profiles of women in Debian http://www.debian.org/women/profile
- Women in Debian and in free software (Magni Onsøien) -
- Statistics about women participation in Debian - https://wiki.debian.org/DebianWomen/Projects/Statistics
- MailingListStats by MetricsGrimoire - http://metricsgrimoire.github.io/MailingListStats/
- The gender.c program by Jörj Michael - http://www.autohotkey.com/board/topic/20260-gender-verification-by-forename-cmd-line-tool-db/
- The US name Database - http://www.census.gov/genealogy/www/data/1990surnames/names_files.html