The goals of this MSR data paper are following:

1. Curated data: To offer a curated data set with data from over 2,000 FLOSS contributors.

2. Combination of data: To present a case study, the challenges and issues of an “augmented” use of the data together with public data from other sources.

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

Combining data: Case study

• 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.
• To automatically infer gender for StackOverflow users, we used a previously-validated name-based gender resolution tool. The tool (https://github.com/tue-mdse/genderComputer) tries to infer a person’s gender based on their name and, if available, their location.
• The samples are composed of 1,476 FLOSS survey respondents out of a total 2,332,406 total MD5 hashes gathered.
• 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.
• Combining data provides with an augmented view of the matter of study.

Combining data: Case study Results

• 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.

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

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.95</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>Recall</td>
<td>0.54</td>
<td>0.39</td>
<td>0.47</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.69</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>MCC</td>
<td>0.96</td>
<td>0.62</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Combining data: Case study Results

• 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.

Methodology

The survey methodology of the FLOSS 2013 survey has been the same as the one of the original FLOSS: an open web-based survey, where participants are self-selected. The 58 questions can be classified into following areas:

• 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, sum money with FLOSS)
• Technology (favorite editor, programming languages)
• Economic and community rewards (job opportunities, expectations from other developers, challenges)

Relevance

• 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.

References