4th International Workshop on Public Data about Software Development

Jesús M. González-Barahona[1], Megan Squire[2], and Daniel Izquierdo-Cortázar[1]

[1] GSyC/LibreSoft, Universidad Rey Juan Carlos
Madrid, Spain
{jgb,dizquierdo}@gsyc.es
[2] Elon University, Department of Computing Sciences
North Carolina, USA
megan@elon.edu

1 Introduction

Libre (free, open source) projects offer publicly available data sources. The research community is starting to produce, use and exchange large data sets of information. These data sets have to be retrieved, purged, described, and can be published for public consumption by other groups. Their availability allows for the decoupling of research activities, the reproducibility of research results, and even the collaboration (and competition) in the analysis of data.

This activity is frequently presented at workshops and conferences, but since the focus of these conferences is not specific to the use of public data, discussions of techniques and experiences are not as deep and fruitful as they could be. This workshop is once again (for the fourth year in a row) such a place. We will host discussions specifically about these sorts of public data sets about software development, how they are retrieved, how they can be analyzed and mined, how they can be exchanged and extended.

2 Main Goals

The goal of this workshop is to foster the analysis of public available data sources about software development and the exchange of data between different research groups. Three different kinds of studies are required (although other related studies could also be considered):

- Analysis of specific projects (provided by the organizers, see below). The analysis should show a methodology to explore the projects, but also it should show explanations to "odd" things that could appear in the data set. For instance, a company-driven project may show different behavior than a community-driven project. The study should be in the field of software engineering, economics, sociology, human resources, and others.
- Retrieval process and exchange formats of public available data collections about software development. The data collections presented should be publicly available,

C. Boldyreff et al. (Eds.): OSS 2009, IFIP AICT 299, pp. 351–352, 2009.

based themselves on public data (so that other groups could reproduce the data collection process), and be related to the field of software development. This includes, but it is not limited to, data from source control systems, bug tracking systems, mailing lists, websites, source and binary code, quality assurance systems, etc.
- Data mining activities and new retrieval tools. Working with a huge quantity of data invites complexity in storage and analysis. Data mining techniques are welcome in this section. This analysis should show a methodology to explore the data and explanations about the whole process. Cross-analysis of datasets is more than welcome. Also, new tools developed to obtain data from several data sources, such as forums, wikis, bug tracking systems and others fit perfectly here.

3 Detailed Description of Data Sources

Two specific issues will be considered together with the development of new data mining tools:

- Data collections analysis about FLOSS development: specifically FLOSSMole and FLOSSMetrics. These collections, already available to any researcher, are offered for analysis by third parties. The studies submitted should detail how they have been used, which part of the information has been considered, how they have been validated or filtered and/or post-processed (if that is the case). The description should be detailed enough to let any other research group reproduce the study.
- Studies about the data retrieval and preparation for public consumption of other data sets in the same realm, which could be proposed for analysis in future editions of the workshop.

3.1 FLOSSMole

FLOSSMole[1] (formerly OSSmole) is a set of tools for gathering data about the development of FLOSS projects. It also publishes the resulting analysis about those projects, and accepts data donations from other research groups. It offers researchers an extensive set of data gathered from the SourceForge development platform and the Freshmeat announcement systems, as well as Rubyforge, Objectweb, Free Software Foundation, and Debian.

3.2 FLOSSMetrics

At the end of FLOSSMetrics[2] a huge database with data from thousands of projects will be available. Nowadays, the project is already working on retrieving data, with information already available for more than 2,000 projects (focused on CVS and SVN repositories, but also mailing lists and issue tracking systems). These results are publicly available at http://melquiades.flossmetrics.org

[1] `http://ossmole.sourceforge.net`
[2] `http://flossmetrics.org`