WoPDaSD 2010: 5th Workshop on Public Data about Software Development

Jesus M. Gonzalez-Barahona[1], Megan Squire[2], and Daniel Izquierdo-Cortazar[1]

[1] GSyC/LibreSoft, Universidad Rey Juan Carlos, Mostoles, Madrid
{jgb,dizquierdo}@libresoft.es
[2] Elon University, North Carolina, The USA
msquire@elon.edu

1 Introduction

Projects such as FLOSSmole and FLOSSMetrics are compiling huge quantities of data about libre (free, open source) software development. The availability of these data in formats suitable for analysis by third parties are enabling researchers to focus on the study of the data, and not on data retrieval activities. This is fortunate, since data retrieval from software development repositories is becoming more and more complex, especially when reliable and detailed information from many projects is needed.

The use for research purposes of this kind of data compiled by teams external to the researcher is posing new problems. Annotation of data, exchange formats, traceability and privacy issues, are becoming issues to be addressed. In addition, working with FLOSS projects to easy obtaining their data, and showing them how that can benefit their activities is also of increasing importance.

Despite these open issues, the use of these open datasets is enabling researchers in many ways: reproduction of results is easier; massive analysis (based on data from hundreds or even thousands of projects) is possible; quick obtaining of results is simplified; availability of data for research communities with little experience in retrieving data from software repositories.

Studies and research results based on this kind of dataset have already been presented in workshops, conferences and journals, but rarely the focus is on how to benefit from the datasets, or on the problems derived from their use. In addition, the details of how to use the datasets for different purposes, or specific results from their analysis, are not published elsewhere.

This workshop is once again (for the fifth year in a row) a place to discuss all these topics, and to present research results developed with these ideas in mind: how these large datasets about FLOSS software development are retrieved, how can they be analyzed and mined, how they can be published, exchanged and extended, which lessons are we learning from their use, and which results are being obtained from their analysis.

2 Goals

The goal of this workshop is to foster the production and analysis of publicly available data sources about software development and the exchange of data between different research groups. The workshop is aimed at the following kinds of studies (although other related studies could also be considered):

- Results based on the analysis of large datasets about software development.This refers mainly to research conducted on FLOSSmole or FLOSSMetrics data, but also on other similar open source datasets. 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 can show different behavior than a community-driven project. The study can be in the field of software engineering, economics, sociology, human resources, and others.
- Retrieval process and exchange formats of publicly available data collections about software development. The data collections presented should be publicly available, 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 is not limited to, data from source code control systems, but tracking systems, mailing lists, websites, source and binary code, quality assurance systems, etc. Although any kind of data collection can be considered, those including information about a large number of projects will be considered especially appropriate.
- 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, provided that papers include some conclusions about a specific set of projects. Again, this analysis should show a methodology to explore the data and explanations about the whole process. Cross-analysis of datasets, and specially of those provided by the organizers (FLOSSMole and FLOSSMetrics databases) is especially welcome. Also, new tools developed to obtain data from several data sources, such as forums, wikis, bug tracking systems and others fit perfectly here.
- Usage of public datasets about software development by new research communities, which until now did little empirical research in this area because they lacked the expertise needed to retrieve information directly from the repositories, but are now empowered by the availability of these datasets. Research results produced by these communities, cases of use, problems found, etc. are possible contributions to the workshop.