2nd International Workshop on Public Data about Software Development (WoPDaSD 2007)

Jesus M. Gonzalez-Barahona, Megan Conklin, Gregorio Robles
http://libresoft.urjc.es/Activities/WoPDaSD2007

Abstract. Exchange of detailed data about software development between research teams, and specifically about data available from public repositories of libre (free, open source) software projects is becoming more and more common. This workshop will explore the benefits and problems of such exchange, and the steps needed to foster it. As a case example of data exchange, the workshop organizers suggest two large datasets to be analyzed by participants.

Introduction

In the latest years, and specially thanks to the huge availability of data about software development that can be obtained from libre (free, open source) projects, 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 (some groups can focus on data retrieval and preliminary analysis, which others can devote to more in-depth analysis without bothering with data retrieval), the reproducibility of research results, and even the collaboration (and competition) in the analysis of data.

All this activity is being presented in several workshops and conferences, but a single place to exchange experiences does not exist yet. We propose this workshop as such a place, where researchers in the field can discuss specifically about this kind of data sets, how they are retrieved, how can they be analyzed and mined, how they can be exchanged and complemented, etc.

Please use the following format when citing this chapter:

Gonzalez-Barahona, J.M., Conklin, M. and Robles, G., 2007, in IFIP International Federation for Information Processing, Volume 234, Open Source Development, Adoption and Innovation, eds. J. Feller, Fitzgerald, B., Scacchi, W., Sillitti, A., (Boston: Springer), pp. 381–383.

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.

The workshop is aimed specifically at two different target studies:

- Analysis of some data collections about software development (provided by the organizers, see below). The analysis should show a methodology for exploring any of those data sets (or better, to relate both) searching for some specific result in the area of software development, and its applications to the actual data sets. The study can 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, 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 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 amount of projects will be considered especially appropriate.

The target audience is composed by the research groups interested in empirical software engineering and quantitative studies of the software development processes and methods. This includes not only software engineers, but also researchers from other fields that might use the data for economic, social and other studies.

Detailed Description

Following the goals described above, the workshop will accept papers about two specific issues:

- Analysis of two data collections about libre software development: FLOSSMole and CVSAnaly-SF. These collections, already available to any researcher, are offered for the analysis. 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 data sets in the same realm, which could be proposed for analysis in future editions of the workshop.

FLOSSMole

FLOSSMole (formerly OSSmole) is a set of tools for gathering data (metrics) about the development of free/libre/open source projects. The FLOSSMole project also publishes the resulting analysis about FLOSS projects, and accepts data donations from other research groups. It offers this workshop a complete set of data gathered from the SourceForge development platform and the Freshmeat announcement systems. More information can be obtained from http://ossmole.sourceforge.net.

CVSAnalySF

CVSAnalY is a tool created by the Libre Software Engineering Group at the Universidad Rey Juan Carlos that extracts statistical information out of CVS (and recently Subversion) repository logs and transforms it in database SQL formats. It has been used to retrieve information for all projects that have an active CVS system at SourceForge. This data set is publicly offered to be analyzed in this workshop. More information can be obtained from http://libresoft.urjc.es/Data.

Challenge

This edition an specific challenge is proposed to contributors, in addition to regular papers. The topic of the challenge is “data visualization”, and will consist on papers about visualization of the data in any of the datasets offered (FLOSSMole, CVSAnaly-SF, or both). The text in the paper should explain the visualization technique used, and its possible applications. The images in the paper should be the visualization images themselves, or snapshots of them. Visualization techniques that help to answer interesting questions, to better understand the data, or to find relationships in it (including relating data in both datasets) are encouraged.