FLOSSMetrics: Free / Libre / Open Source Software Metrics

Israel Herraiz, Daniel Izquierdo-Cortazar, Francisco Rivas-Hernandez,
Jesús Gonzalez-Barahona, Gregorio Robles, Santiago Dueñas-Domínguez,
Carlos Garcia-Campos, Juan Francisco Gato, Liliana Tovar
GSyC/LibreSoft
Universidad Rey Juan Carlos (Madrid, Spain)
{herraiz, dizquierdo, frivas, jgb, grex, sduenas, carlosgc, jfcogato, lilitovar}@gsyc.urjc.es

Abstract

This paper presents FLOSSMetrics, a European Commission 6th Framework Programme-funded research project. The main objective of FLOSSMETRICS is to construct, publish and analyse a large scale database with information and metrics about libre software development coming from several thousands of software projects, using existing methodologies, and tools already developed. The project also provides a public platform for validation and industrial exploitation of results. The project is in its final stage, and some results and databases are already available, as is shown in this paper.

1 Name and acronym

Free/Libre/Open Source Software Metrics and Benchmarking Study - FLOSSMETRICS

2 Source and amount of funding

FLOSSMetrics is funded by the European Commission 6th Framework Programme under contract #FP6-033982. The total budget of the project is 585,100 Euros, with a project funding of 583,800 coming from the European Commission.

3 List of participants

The FLOSSMetrics consortium consists of universities and companies from different European countries, with a role that is detailed below:

- *Universidad Rey Juan Carlos*, Spain. Coordinator of the project, and research partner in charge of the development of the retrieval system of the project.
- *University of Maastricht*, The Netherlands. Another research partner, whose main work is the development of the exploitation plan by large firms and small and medium enterprises (SMEs)
- *Wirtschaftsuniversitaet Wien*, Austria. Research partner in charge of the visualization of the results produced by the project.
- *Aristotle University of Thessaloniki*, Greece. Research partner whose main work is the identification and evaluation of the different data sources that feed the retrieval system.
- *Conecta s.r.l.*, Italy. An industrial partner that is helping to identify the opportunities of exploitation by SMEs.
- *ZEA Partners*, Belgium. Another industrial partner, that is helping in the exploitation of results by both SMEs and large firms (ZEA Partners is a consortium of SMEs, so it features the behavior both of a SME and of a larger firm).
- *Philips Medical Systems PMS Nederland B.V.*, The Netherlands. Industrial partner, that explores the possibilities of exploitation of the project within large firms.

This consortium has also enabled a widespread dissemination of results, in the scope of the academia, SME, large firms, and the FLOSS community (with participation in events such as FOSDEM[1] and Linuxtag[2]).

4 Status and duration of the project

4.1 Status and outcome

FLOSSMetrics is entering its final stage, and it has already available some information about a sample of

[1] http://www.fosdem.org
[2] http://www.linuxtag.org

1,000 FLOSS (free / libre / open source software) projects (approximately, because this number is progressively increasing until reaching the goal of the project of 5,000 projects). This information comes from different data sources, mainly: source code management repositories, mailing lists archives, bug tracking systems and source code.

This information is gathered in an automatic way, thanks to the *retrieval system*, developed under the scope of the project. This system is publicly available [3], and it integrates some other mining tools developed by the consortium and some other third parties:

- **CVSAnalY**. It extracts information from CVS, Subversion or Git repository logs and transforms it in a database SQL format.
- **Mailing List Stats**. Tool for mapping mbox files of any mailing list to a database.
- **Bicho**. It stores information from a given bug tracking system (BTS) to a database format. So far, it only works with the SourceForge.net's BTS.
- **Libresoft's CMetrics package**. It provides complexity metrics for source code files written in C.
- **SLOCCount**. This tool, developed by David A. Wheeler[4], counts physical source lines of code (SLOC) in large software systems and it provides a basic COCOMO model results.
- **CCCC**. It provides metrics for C, C++ and Java files.
- **PyMetrics**. It provides metrics for Python files.
- **PerlMetrics**. It provides metrics for Perl files.

The results of the retrieval system are offered through a web interface, called *Melquiades*[5]. So far, four kind of repository metrics are offered: source code management information, mailing lists, code metrics (only for files written in C) and bug tracking system information.

We recommend to visit the Melquiades web interface, in order to check the kind of results that are provided by the project. In particular, we strongly recommend to explore the provided databases to researchers interested in the study of FLOSS, because it is a publicly available and common dataset that makes it easier to verify the results of those studies. We also strongly recommend to any FLOSS developer to suggest the inclusion of any project that may be missing from the list of FLOSS projects analyzed by FLOSSMetrics.

[3] http://forge.morfeo-project.org/projects/libresoft-tools/

[4] See http://www.dwheeler.com/sloccount

[5] http://melquiades.flossmetrics.org/

4.2 Duration

The duration of the project is from September 2006 to September 2009 (36 months).

5 Goals

The main goals of the project are the following:

- Identify and evaluate sources of data and develop a comprehensive database structure, built upon the results of CALIBRE[6], another research project funded by the Framework Programme.
- Integrate already available tools to extract and process such data into a complete platform.
- Build and maintain an updated empirical database applying extraction tools to thousands of open source projects.
- Develop visualisation methods and analytical studies, especially relating to benchmarking, identification of best practices, measuring and predicting success and failure of projects, productivity measurement, simulation and cost/effort estimation.
- Disseminate the results, including data, methods and software.
- Provide for exploitation of the results by producing an exploitation plan, validated with the project participants from industry especially from an SME perspective.

More in detail, the main goal of FLOSSMetrics is to build, publish and analyse a large scale of projects, retrieving information and metrics about libre software development. There are dozens of developed tools, both, among partners and libre software ones. Some of these tools are being integrated in the FLOSSMetrics platform, which provides a public web site for validation and industrial exploitation of results.

Focusing on the software development itself, this project aims to provide information about the actors (developers), the artifacts, source code and processes. At the end of this project, thousands of libre software projects will be analysed and in the long term, the studies made possible by the data obtained, will allow the identification of techniques and procedures for estimating the future of a project with a certain probability. Also, an additional goal is to retrieve data to measure the productivity rate, with the aim of getting accurate indicators for libre software projects.

[6] See http://calibre.ie for the results of this project, that finished in 2006.

We expect that the impact of the project will be large in the libre software development world, but also in the research and industrial world. FLOSSMetrics is expected to produce the biggest and most complete and detailed view of the current skyline of libre software development. Both, providing a statistical point of view and providing historical data from all of them. In this way, this project tries to better understand the evolution of libre software projects offering data of them in each step of their life. In other words, one of our goals is to provide quantitative data (data from thousands of projects and dozens points of view) and qualitative data, from the point of view of software engineering.

Focusing on the software engineering field, this database will allow researchers to test several assumptions and models currently discussed, including non-common development paradigms, like aspect oriented development.

Finally, centralizing efforts in a database like the one FLOSSMetrics is building, we try to facilitate the process of traceability. In most of the cases the scripts, tools or regular expressions used for specific analysis are not public and even most of the works cannot be reproduced again. Using this database, everybody is able to use the offered data and to trace the results and even do the experiments from scratch.

6 Relevance of the Project

In the Software Engineering (SE) field, there is a concern shared by some researchers about the lack of empirical research and experimentation [2, 3, 4], and about the limited impact of SE studies in industry settings [1, 5, 6],

In particular, the FLOSS community is a great opportunity to enhance and stimulate empirical research in the scope of SE. There are thousands of projects in that community, and most of them provide their repositories for anyone with any purpose.

However, in spite of this high availability of data, there are some problems inherent to the study of FLOSS. First of all, to certain extent, these repositories of data are heterogeneous, which makes it difficult to extend the studies to a broad sample of FLOSS projects. Secondly, the retrieval and analysis of those heterogeneous data sources require to develop software tools, that most of the times are ad-hoc and therefore very difficult to reuse by third parties, which is a barrier to the verifiability of the results and the methodologies of the studies. Thirdly, this kind of studies demands a lot of resources of the FLOSS projects (CPU usage by the servers, bandwidth), and those projects do not perceive immediate benefits out of those studies; these two facts provoke a climate of hostility in the FLOSS projects towards research activities.

These problems can be overcome by using common and publicly available datasets, such as those provided by FLOSSMetrics. The first problem is solved thanks to the retrieval system previously mentioned. FLOSSMetrics has developed a data model for that system that aggregates results coming from different versions of the same kind of repositories (for instance, Git and Subversion for source code management systems), which helps to lower the heterogeneity of the data sources. Moreover, because the datasets are publicly available and regularly updated, researchers have not worry about the development of software tools for the retrieval of the data. This also makes it possible to verify the studies by third parties, because the same datasets are shared by all the researchers. Furthermore, it allows to easily extend these studies to a large amount of case studies, as the number of FLOSS projects stored in the FLOSSMetrics databases is in the order of thousands. The processes followed to retrieve, store and distribute those databases also ensure the traceability of the results. Finally, because the datasets are obtained directly from the FLOSS projects only once, this lowers the stress that the FLOSS projects' systems suffer. In addition to that, FLOSSMetrics is providing regular reports about all the projects stored in its databases, in the hope of contributing back to those projects. This is helping to establish a favorable climate towards research activities within the FLOSS community. The spread of the results of FLOSSMetrics in events such as FOSDEM and Linuxtag also helps to achieve that goal.

Regarding the impact of empirical studies in industrial settings, some of the partners of the FLOSSMetrics consortium comes from the industry, both at the SME and the large firms level. The main goal is to help to spread the benefits of using FLOSS within companies. This may look at a questionable affirmation, and actually it is. But precisely the goal of FLOSSMetrics is to backup that argument with the facts found in the different analysis and reports done both by the consortium and by third parties using the FLOSSMetrics datasets. As an initial result, FLOSSMetrics has produced the *FLOSS: a guide for SMEs* guide[7], that helps to enlighten the adoption process of FLOSS in the IT departments of SMEs. In the case of large firms, the participation of Philips Medical Systems is helping to show some experience cases in the adoption and development of FLOSS in the software secondary sector, this is, in large firms that do not develop nor sell software as their main activity, but heavily rely on it for their businesses.

In summary, FLOSSMetrics is gathering information about thousands of libre software projects, and making those datasets available for third parties. This is helping to make it easier to study SE in general and FLOSS in particular from a empirical point of view. Furthermore, those datasets are helping to provide facts and figures to help in the spread and adoption of FLOSS both in SMEs and large companies.

[7] Available at http://guide.conecta.it/

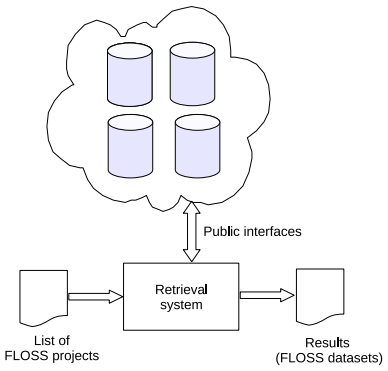

Figure 1. Approach used by FLOSSMetrics.

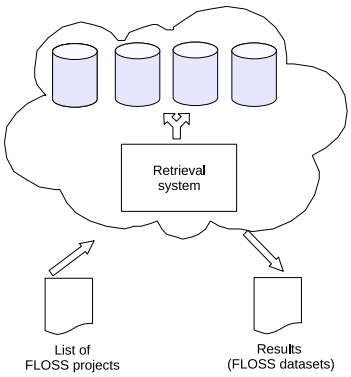

Figure 2. Suggested approach for Research-Friendly.

7 Future of the project

The FLOSSMetrics project is expected to finish on September 2009. However, this project is a milestone in a longer term roadmap that we denominate *ResearchFriendly*.

We have already highlighted that there is a lack of empirical research in the scope of SE, and that the impact of empirical studies in industry settings is still low. We have also mentioned the problems of performing empirical studies of FLOSS, and how an indiscriminate retrieval of data requests to the FLOSS projects' repositories may generate a climate of hostility towards research activities.

FLOSSMetrics is addressing those problems, by providing publicly available datasets, reports for the FLOSS projects under study, and by collaborating in FLOSS events where the goals and results of the projects are discussed together by researchers and FLOSS developers.

However, there are some difficulties in the approach used by FLOSSMetrics, shown in figure 1. It uses the public interfaces provided the FLOSS projects, which usually requires more resources than doing the same operations internally in the repositories, and sometimes implies that some information is missed or has to be reconstructed by using heuristics.

For the Research Friendly initiative, we propose the data to be directly provided by the FLOSS projects, as illustrated in figure 2. This would lower the stress in the repositories, and all the problems associated to the retrieval of information through public interfaces (like for instance, a ban when a server gets overloaded). This of course would need of the intimate collaboration of FLOSS projects, by installing additional plugins in their repositories. To achieve this objective, researchers and FLOSS projects should collaborate closer. The participation in FLOSS events is a first step towards this end, although the commitment of FLOSS projects will be never achieved unless the benefits of this kind of research are clearly remarked.

Summarizing, the future of FLOSSMetrics is to try to commit FLOSS projects in the retrieval and distribution of information, by means of providing meaningful and useful results for those projects.

8 Contact and acknowledgements

The authors of this report would like to thank to all the members of the FLOSSMetrics consortium for their collaboration in the elaboration of this report.

The project coordinator can be reached at: Jesús M. González-Barahona, Universidad Rey Juan Carlos, email: jgb@gsyc.urjc.es.

More information about the project may be found at http://flossmetrics.org

References

[1] D. Budgen and B. Kitchenham. Realising evidence-based Software Engineering: a report from the workshop held at ICSE 2005. *ACM SIGSOFT Software Engineering Notes*, 30(5):1–5, 2005.

[2] T. Dyba, B. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *IEEE Software*, 22(1):58–65, 2005.

[3] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, pages 721–734, 2002.

[4] L. Pickard, B. Kitchenham, and P. Jones. Combining empirical results in software engineering. *Information and Software Technology*, 40(14):811–821, 1998.

[5] W. Tichy. Should computer scientists experiment more? *Computer*, pages 32–40, 1998.

[6] M. V. Zelkowitz, D. R. Wallace, and D. W. Binkley. *Lecture notes on empirical software engineering*, chapter Experimental validation of new software technology, pages 229–263. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2003.