Chapter 1.9
FLOSSmole:
A Collaborative Repository for FLOSS Research Data and Analyses

James Howison
Syracuse University, USA

Megan Conklin
Elon University, USA

Kevin Crowston
Syracuse University, USA

ABSTRACT

This article introduces and expands on previous work on a collaborative project, called FLOSSmole (formerly OSSmole), designed to gather, share, and store comparable data and analyses of free, libre, and open source software (FLOSS) development for academic research. The project draws on the ongoing collection and analysis efforts of many research groups, reducing duplication, and promoting compatibility both across sources of FLOSS data and across research groups and analyses. The article outlines current difficulties with the current typical quantitative FLOSS research process and uses these to develop requirements and presents the design of the system.

INTRODUCTION

This article introduces a collaborative project called FLOSSmole,[1] designed to gather, share, and store comparable data and analyses of free and open source software development for academic research. The project draws on the ongoing collection and analysis efforts of many research groups. Our intent in developing FLOSSmole is to reduce duplication, and to promote compatibility both across sources of FLOSS data and across research groups and analyses.

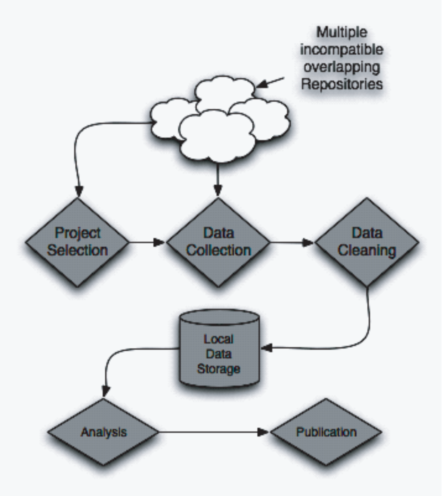
Creating a collaborative data and analysis repository for research on FLOSS is important because research should be as reproducible, extendable, and comparable as possible. Research with these characteristics creates the opportunity to employ meta-analyses, exploiting the diversity of existing research by comparing and contrasting results to expand our knowledge. Unfortunately, the current typical FLOSS research project proceeds in a way that does not necessarily achieve these goals. These goals require detailed communal knowledge of the many choices made throughout a research project. Traditional publication prioritizes results, but masks or discards much of the information needed to understand and exploit the differences in our data collection and analysis methodologies. FLOSSmole was originally designed to provide resources and support to academics seeking to prepare the next generation of FLOSS research. Since its inception, FLOSSmole has also been a valuable resource for nonacademics who are also seeking good data about development practices in the open source software industry.

BACKGROUND OF PROBLEM

Obtaining data on FLOSS projects is both easy and difficult. It is easy because FLOSS development utilizes computer-mediated communications heavily for both development team interactions and for storing artifacts such as code and documentation. This way of developing software leaves a freely available and, in theory at least, highly accessible trail of data upon which many academics have built interesting analyses about optimal organization of development teams, economics of building software in the commons, and the like. Yet, despite this presumed plethora of data, researchers often face significant practical challenges in using this data to construct a collaborative and deliberative research discourse. In Figure 1, we outline the research process we believe is followed in much of the quantitative literature on FLOSS.

The first step in collecting online FLOSS data is selecting which projects and which attributes to study, two techniques often used in estimation and selection are census and sampling. (Case studies are also used but these will not be discussed in this article.)

Figure 1. The typical quantitative FLOSS research process (notice its noncyclical and noncollaborative nature)



Conducting a census means to examine all cases of a phenomena, taking the measures of interest to build up an entire accurate picture. Taking a census is difficult in FLOSS for a number of reasons. First, it is hard to know how many FLOSS projects there are "out there," and it is hard to know which projects should actually be included. For example, are corporate-sponsored projects part of the phenomenon or not? Do single-person projects count? What about school projects?

Second, the projects themselves, and the records they leave, are scattered across a surprisingly large number of locations. It is true that many are located in the major general repositories, such as Sourceforge[2] and GNU Savannah.[3] It is also true, however, that there are a number of other repositories of varying sizes and focuses (e.g., CodeHaus,[4] CPAN[5]), and that many projects, including the well-known and much-studied Apache and Linux projects, prefer to use their own repositories and their own tools. This diversity of location effectively hides significant portions of the FLOSS world from attempts at census. Even if a full listing of projects and their locations