Chapter 3.5
Tools and Datasets for Mining Libre Software Repositories

Gregorio Robles
Universidad Rey Juan Carlos, Spain

Jesús González-Barahona
Universidad Rey Juan Carlos, Spain

Daniel Izquierdo-Cortazar
Universidad Rey Juan Carlos, Spain

Israel Herraiz
Universidad Alfonso X el Sabi, Spain

ABSTRACT

Thanks to the open nature of libre (free, open source) software projects, researchers have gained access to a rich set of data related to various aspects of software development. Although it is usually publicly available on the Internet, obtaining and analyzing the data in a convenient way is not an easy task, and many considerations have to be taken into account. In this chapter we introduce the most relevant data sources that can be found in libre software projects and that are commonly studied by scholars: source code releases, source code management systems, mailing lists and issue (bug) tracking systems. The chapter also provides some advice on the problems that can be found when retrieving and preparing the data sources for a later analysis, as well as information about the tools and datasets that support these tasks.

1. INTRODUCTION

In libre software[1] projects communication and organization are heavily dependent on the use of telematic means. Face-to-face communication is rare, and Internet-based tools are the most common means for a developer to interact with the code and with other developers.

Fortunately for researchers, the data produced by those interactions is usually stored and offered publicly over the Internet. The repositories for these data contain information valuable to understand the development process, and can be

DOI: 10.4018/978-1-61350-456-7.ch3.5

analyzed in combination with the most classical data source: source code. In addition, the ability of having detailed information from the past (since it is usually archived for long periods of time) offers the possibility of performing also longitudinal and evolutionary analysis.

Research groups worldwide have already taken benefit from the availability of such a rich amount of data sources in the last years. Nonetheless, the access, retrieval and fact extraction is by no means a simple task and many considerations and details have to be taken into account to successfully retrieve and mine the data sources.

This chapter offers a detailed description of the most common data sources that can generally be found for libre software projects on the Internet, and of the data that can be found in them: source code releases, source code management systems (in the following, SCM), mailing lists archives, and issue or bug tracking system (in the following, BTS). In addition, we present some tools and datasets that might help researchers in their data retrieval and analysis tasks.

Mining and analyzing these data sources offer an ample amount of possibilities that surpass or complement other data-acquiring methodologies such as surveys, interviews or experiments. The amount of data that can be obtained, in a detailed way and in many cases for the whole lifetime of a software project, gives a precise description of the history of a project (Bauer and Pizka, 2003). In this sense, we have access to the activities (the what), the points in time (the when), the actors (the who) and sometimes even the reason (the why) (Hahsler and Koch, 2005). Compared to surveys, mining these data sources allow to access data for thousands of developers and a wide range of software projects. Most of these efforts can be considered as non-intrusive, as researchers can analyze the projects without interacting with developers, which is friendly to them. But even in a small environment, e.g., when evaluating the impact of software tools in a small team (Atkins et al., 2002), the use of data from one or more of these sources provides additional insight. Furthermore, mining software repositories has many advantages compared to conducting experiments as real-world software projects are taken into consideration (Mockus and Votta, 2000, Graves and Mockus, 1998).

2. FIRST STEPS BEFORE THE ANALYSIS

There are some steps to be walked before the analysis of data from libre software projects can be started. First of all, the relevant data sources have to be identified. After that, the data has to be retrieved from the corresponding data repositories. Only then, the researcher can really start to analyze the data.

It is important to notice that there may be several ways of accessing the same kind of data, depending on the project and how it handles it. There are several different tools and systems that projects use, and they also have different usage conventions. For instance, the use of tags, comments, among others, may differ from one project to another, and can be of paramount importance to tell bugs appart from new feature requests in a BTS. The complexity and feasibility of both identification and retrieval depend, therefore, of the project. Figure 1 shows a diagram with all the steps that have to be accomplished for any source considered in the studies.

Figure 1. Whole process: from identification of the data sources to analysis of the data