Chapter 3.6
Tools for the Study of the Usual Data Sources found in Libre Software Projects

Gregorio Robles
Universidad Rey Juan Carlos, Spain

Jesús M. González-Barahona
Universidad Rey Juan Carlos, Spain

Daniel Izquierdo-Cortazar
Universidad Rey Juan Carlos, Spain

Benjamin E. Erlandson
Universidad Rey Juan Carlos, Spain

ABSTRACT

Due to the open nature of Free/Libre/Open Source software projects, researchers have gained access to a rich set of development-related information. Although this information is publicly available on the Internet, obtaining and analyzing it in a convenient way is not an easy task and many considerations have to be taken into account. In this paper we present the most important data sources that can be found in libre software projects and that are studied by the research community: source code, source code management systems, mailing lists and bug tracking systems. We will give advice for the problems that can be found when retrieving and preparing the data sources for a posterior analysis, as well as provide information about the tools that support these tasks.

INTRODUCTION

The fact that communication and organization are heavily tied in libre software[1] projects to the use of telematic means and that these interactions are, in general, stored and publicly offered over the Internet makes the number of data sources where development information can be extracted from grow beyond source code. In addition, the ability of having memory (as data from activities in the past can be obtained) offers the possibility of performing longitudinal analysis as well. Research groups from all around the world 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 have to be considered to successfully mine the data sources.

This chapter is probably the first attempt to have a detailed description of the most common data sources that can generally be found for libre software projects on the Internet and the data that can be found in them. In addition, we present some available tools that might help researchers obtaining and partially analyzing the described data sources. These data sources comprise **source code**, **source code management** (in the following, SCM), **mailing lists archives**, and **bug tracking system** (in the following, BTS).

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 allows 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. 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).

The structure of this chapter is as follows: the next section handles the identification of the data sources as well as its retrieval process. Next, various analysis on source code are introduced (hierarchy, file discrimination, analysis of *traditional* source code files, analysis of the rest of files (such as documentation, multimedia, etc.), and authorship). The fourth section presents how SCM systems can be mined, putting special attention on the CVSAnalY tool and some details to be considered when performing analyses on CVS. The fifth section presents the most common format in which mailing lists are stored (MBOX), while the sixth one is devoted to present the data to be found in a BTS. Finally, the reader can find a short summary of the chapter in the last section.

IDENTIFICATION OF DATA SOURCES AND RETRIEVAL

There are some steps before the analysis of data from libre software projects can be started that should be considered: identification and retrieval. It should be noted that there may be several ways of accessing the data, depending on the projects. This is because of the use of the several development-supporting tools that projects use and of having different usage conventions (for instance, the use of tags, comments, among others, may differ from one project to another). The complexity and feasibility of both activities depend on the data source and on the project. Figure 1 gives a diagram that shows the steps that have to be accomplished for any source considered in our study.

In general terms, the identification of the data source depends mostly on its significance for the software development of a project. Hence, identifying the source code, the SCM system, the mailing lists or the BTS is in no way problematic