

# Getting the global picture

Abstract of OWLS Presentation  
*Oxford Internet Institute, 25-26<sup>th</sup> June 2004*

**Jesus M. Gonzalez-Barahona**

&

**Gregorio Robles-Martinez**

Informatics, Universidad Rey Juan Carlos  
[jgb@gsyc.esct.urjc.es](mailto:jgb@gsyc.esct.urjc.es), [grex@gsyc.esct.urjc.es](mailto:grex@gsyc.esct.urjc.es)

Libre (free, open source) software projects usually maintain large quantities of data available in public repositories, ranging from version control systems to bug tracking databases to web-based discussion forums or mailing list archives. Most of this information (but not all) can be retrieved and analyzed automatically, and several works have been presented in the latest years about how to exploit them from a number of points of view.

Currently, those studies tend to focus on given projects or certain aspects of the information about some projects. The problem of getting the "global picture" of the libre software landscape is still an open issue. Cases of detailed studies of large collections of projects (in the hundreds), crossing information from several different kinds of repositories or analysis of historic patterns in different projects over long periods of time are still rare.

However, those cases are basic to deal with questions like what motivates volunteer developers (which involves tracking the history of individuals through several projects and several roles), how can volunteer work be exchanged by remunerated work (which requires analyzing effort and dedication in many projects with different profiles of involvement of companies) or estimation of the importance of code reuse (which requires the historical analysis of source code along large quantities of projects).

Unfortunately, getting that global picture poses difficult problems: managing huge quantities of data with inconsistencies, missing information, inaccuracies, and different representations.

Our presentations will show first some tools (and results) that allow one to manage the quantity of libre software data publicly available on the Internet. Second, it will try to emphasize which problems are foreseeable in this context, how to deal with them, and what steps could be followed in some of the open questions that have been identified.