Data and analyses sharing to support research on free/libre open source software

A Debate

Brian Fitzgerald[1], Moderator

Proposition	**Opposition**
Evangelia Berdou[2]	Megan Conklin[5]
Kevin Crowston[3]	Stefan Koch[6]
Greg Madey[4]	Walt Scacchi[7]

1 University of Limerick, brian.fitzgerald@ul.ie

2 London School of Economics	5 Elon University
E.Berdou@lse.ac.uk	mconklin@elon.edu
3 Syracuse University	6 Wirtschaftsuniversität Wien
crowston@syr.edu	stefan.koch@wu-wien.ac.at>
4 University of Notre Dame	7 University of California, Irvine
oss@nd.edu	wscacchi@ics.uci.edu

Be it resolved, that the FLOSS research community requires that data and analyses behind FLOSS research publications be made expeditiously available to other researchers.

Research on FLOSS has relied on several different kinds of scientific evidence, such as the archives created by the FLOSS developers, versioned code repositories, mailing list messages and bug and issue tracking repositories [1]. FLOSS teams retain and make public archives of many of their activities as by-products of their open technology-supported collaboration. However, the easy availability of primary data provides a misleading picture of ease of conducting research on FLOSS. Precisely because these data are by-products, they are generally not in a form that is useful for researchers. Instead potentially useful data is locked up in HTML pages, CVS log files, text-only mailing list archives or dumps of website databases. FLOSS research projects, therefore, expend significant energy collecting and re-structuring these archives for their research, which is repetitive and wasteful [2]. Furthermore, different researchers will extract different data at different points in time, take different approaches to processing and cleaning data and make different decisions about analyses, but without all of these decisions being visible, auditable or reproducible. In principle, these latter problems can be addressed by individual researchers better documenting what they have done. However, research publications

typically have restrictions on publication lengths that make complete discussion impossible. Furthermore, published papers are just the tip of the iceberg, and knowing what others have done does not necessarily make it any easier to replicate the results.

In light of these issues, FLOSS research might be greatly facilitated by increased sharing of primary data as well as various stages of data analysis. Such data archives have had some success in facilitating research in other fields, e.g., in biomedical sciences. However, there are numerous problems that must be addressed to make such data sharing feasible and valuable. One of the most important issues is ensuring the appropriate rewards and incentives for sharing. The experience in other fields suggests that if data sharing is an option, it is one that will be exercised by only a few researchers. On the other hand, it is not clear how sharing might be enforced or what the effects of such a mandate might be. The issue seems like one that is ready for a public debate.

Therefore, we will debate the resolution that the F/L/OSS research community requires that data and analyses behind F/L/OSS research publications be made expeditiously available to other researchers. If this resolution gains support from participants at the *International OSS Conference*, then efforts can be made to implement this resolution in the research field.

References

1. D. German and A. Mockus, in *Proceedings of the ICSE 3rd Workshop on Open Source*. (2003).
2. J. Howison, M. Conklin, and K. Crowston, FLOSSmole: A collaborative repository for FLOSS research data and analyses, *International Journal of Information Technology and Web Engineering* **1**(3), 17 (2006).