eSocialScience for Free/Libre Open Source Software researchers

Kevin Crowston, James Howison, Andrea Wiggins

Syracuse University School of Information Studies

crowston@syr.edu

Abstract. This abstract presents a case study of the potential application of eScience tools and practices for the social science research community studying Free/Libre Open Source Software (FLOSS) development practices. We first describe the practice of research on FLOSS to motivate the need for eScience. After outlining suitable public data sources, we describe our initial efforts to introduce eScience tools for FLOSS research, potential obstacles and how the use of such tools might affect the practice of research in this field.

Introduction

Free/Libre Open Source Software (FLOSS) is software developed by a globally distributed volunteer developer base and released under a license allowing further distribution and modification of the source. FLOSS development represents a novel approach to software development, making its work practices of great interest to empirical software engineering researchers. As well, scientists have used FLOSS projects as accessible examples of other phenomena of interest. For example, many study FLOSS teams as examples of virtual teams, as they are dynamic, self-organizing teams comprising distributed professionals, users and others working together in a loosely coupled fashion. Characterized by a rapid and reliable software development process that includes developers from around the world, effective FLOSS development teams somehow profit from the advantages and overcome the disadvantages of distributed work, again making their work practice interesting to study.

To study the work practices of FLOSS developers, researchers draw on repositories of source code, developer interactions and other project data. For example, to study distributed decision-making processes, Heckman et al. (2007) analyzed steps in decision-making episodes embedded in exchanges of email among developers. In the past, FLOSS research often involved redundant data collection by independent research groups, usually through spidering Sourceforge or similar FLOSS development sites (Conklin *et al.*, 2005; Howison *et al.*, 2006a). However, significant progress has been made in the past few years in creating shared data sets held in what have been called Repositories of Repositories (RoRs) (Antoniades *et al.*, 2007). Table 1 summarizes available sources of FLOSS data.

Note though that Table 1 indicates substantial gaps, with plans to fill only some of them. For example, a comprehensive source of mailing lists for projects is not yet available, nor are there plans to archive projects' IRC communications or to document dependencies between projects. Perhaps more significantly, there is a myriad of project-specific information found in the course of research that is not contained in these databases. In addition, there are both social and technical barriers to entry for using and contributing to research data repositories,

	Repository for Research Data[1]:	FLOSS mole	Notre Dame dumps	FLOSS metrics & CVSanalY	Qualoss & SQO-OSS	Source kibitzer
Project Demographics[2]	Basic data	■	■	■		
	Confirmed Locations			■		■
Developer Demographics	Memberships	■	■			■
	Roles	■	■			■
Communication Venues	Mailing lists	▒		▒		
	Forums	▒	■			
	Issue Trackers	▒	■	▒		
	IRC logs					
	Release System	■	■			
Software Venues	SVN/CVS (counts)	■	■	■		
	SVN/CVS full			■		
	Packages produced	■	■	■		
	Releases + Dates	■	■	■		
	Size (LOC, SLOC)			■	▒	■
	Dependencies					
	Complexity Metrics				▒	■
Use and Popularity	Downloads	■	■			
	Pageviews	■	■			
	User ratings	■				
	In Debian	■		■		
	Actual Use[3]	▨				
Sample Collected	Sourceforge	■	■	▨	See Note 5	See Note 4
	Rubyforge	■				
	ObjectWeb	■		▨		
	Savannah (GNU)			See Note 6		
	Debian Distribution	■				
	Apache Foundation			▨		
	GNOME meta-project			▨		
	KDE meta-project				▒	

Key:

□ Not Collected ▒ Planned (or pilot data only)

▨ Partial (selected sub-collection) ■ Present

1. This table excludes services with data not easily available to researchers. Ohloh, for example, was excluded for this reason. The Notre Dame dumps require signing a research usage agreement. Sourcekitbizer was included insofar as it provides public access to data via the FLOSSmole project. FLOSSmetrics includes the earlier sets released by the Libre Software engineering group (CVSanalY and Debian Counts). Qualoss and SQO-OSS are included together for reasons of space: they are separate but collaborating projects.

2. Project Demographics include Names, Descriptions, Founding date, Intended Audience, Operating System/environment, License, Programming language, Maturity/Status and Donors. Projects are often hosted on more than one service, or provide their own services (such as Trac, SVN etc) Confirmed Locations refers to a human effort to identify the locations actually used by each project.

3. Actual use as measured, for example, by the Debian Popularity contest, for which an agent installed by some Debian users reports frequency of package use.

4. Sourcekibitzer samples only Java projects and accepts user contributions (specify project, SCM location, homepage)

5. Qualoss intends to implement their measures on 50 projects, currently there are 5 available as pilot data. SQO-OSS works closely with the KDE meta-project.

6. FLOSSmetrics aims to have validated data for 3,000 projects and currently has partial data available (primarily CVSanalY) for 100.

URLS: FLOSSmole: ossmole.sf.net, Notre Dame: nd.edu/~oss/, FLOSSmetrics: data.flossmetrics.com, CVSanalY: libresoft.es/Results/CVSAnalY_SF, Qualoss: qualoss.org, SQO-OSS: www.sqo-oss.eu, Sourcekibitzer: sourcekibitzer.org. Thanks to Jesus González-Barahona, Gregorio Robles and Megan Conklin for assistance in preparing this table.

Table 1. Publicly available FLOSS research data.

not the least of which is the federated nature of FLOSS data repositories. As there are currently no adopted FLOSS ontologies or metadata standards to ensure portability and interoperability, researchers who wish to use data from multiple repositories must tailor their efforts to the specifics of each data source. As well, because there are no established community data standards, researchers donating data to a repository is still relatively rare.

Furthermore, while much raw FLOSS data are easily available, most are by-products of the teams' work rather than created for scientific use (as can be seen in Table 1). Considerable effort is required to process such data into scientifically meaningful measures of theoretically interesting concepts. For example, the Heckman et al. (2007) study mentioned above required labour-intensive content analysis of messages to find instances of decision triggers, identification and assessment of alternatives, and choice. More generally, most researchers are interested in team performance, but this concept can be operationalized in many different ways (Crowston *et al.*, 2006) and there is not much commonality among studies in the approaches selected nor explicit discussion of the implication of the choice. In summary, while raw data are sometimes shared, there is nearly no sharing of processed data that more directly measures theoretical concepts or of analysis approaches for deriving such measures. Researchers have generally stuck to their preferred data manipulation and statistical analysis tools, conducting in-house development where required. The few researchers who have made their analysis components and workflows available (e.g., Howison *et al.*, 2006b; Robles *et al.*, 2005) have done so simply by placing their tools on their project websites.

Scientific workflow tools

This situation described above is not unfamiliar to practitioners of eScience; fields such as bioinformatics have addressed similar challenges through advances in eScience. This section briefly highlights workflow tools and describes efforts to encourage their use in the FLOSS research community. Scientific workflow tools support high-level programming that binds together data sources and analysis procedures (e.g., Taverna, http://taverna.sourceforge.net and Kepler, http://kepler-project.org). Steps in the analysis are performed by modular components with multiple input and output ports through which the components are linked. These workflows can be represented as a flow diagram (see Figure 1 below) and saved in a single file, which permits sharing and repetition of the analysis. As with most programming environments, much of these tools' utility derives from the included library of components, whether local (e.g., Java or R) or remote (e.g., SOAP web-services). Taverna's developers have also created a social networking site, MyExperiment.org, to encourage sharing of workflows, offering a venue for peer support among researchers.

As a proof of concept of the applicability of eScience ideas to FLOSS research, the authors (with Megan Conklin) have received US National Science Foundation funding to (among other things) replicate with Taverna workflows a small number of studies from the research literature. The current candidate studies for replication are shown in Table 2 and an example workflow in Figure 1. These studies were chosen because they draw on the large data sets described above (FLOSSMole, CVSanalY and the Notre Dame dumps) and span a range of research questions and approaches, with a focus of social networks. The workflows are developed by reading the methods sections of the papers to identify data and analysis approaches and building a workflow to perform the analysis. The workflows, together with original research drawing on the tools, are being disseminated through NSF workshops and FLOSS research conferences.

Our replication effort has several benefits. First are foremost, it provides evidence to the community of the applicability of the tools. Second, by replicating the studies using shared

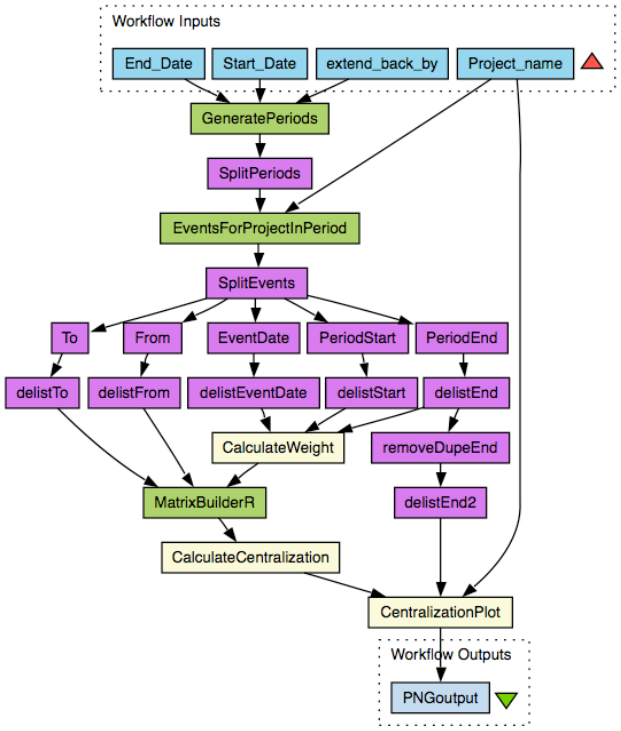

Figure 1. An example workflow that creates a plot of communications centralization over time for a FLOSS project (Howison et al., 2006b). Blue boxes are input and outputs; green are remote processes; purple are built-in processes; and tan are local R scripts.

Table 2. Studies chosen for replication.

Study	**Description**
(Christley & Madey, 2007)	Analyzes how social positions of activity on SourceForge change over time.
(Conklin, 2004)	Examines if developers joining projects create a scale-free network.
(Howison et al., 2006b)	Examines social network of project communications over time.
(Robles et al., 2005)	Examines growth rate of software.
(English & Schweik, 2007)	Classifies FLOSS projects based on metrics for success versus abandonment and stage of project growth.

data sources and analysis workflows, we can extend the original analyses, e.g., by testing alternate choices of variables used to determine project success or applying the analysis to additional data covering larger periods of time or more projects. Third, the effort will also provide a set of local and SOAP components that can be reused in other workflows. In developing the workflows, we encountered issues that illustrate the value of the approach. For example, one paper developed a classification of projects based on their size, rate of growth and other parameters. However, attempts to develop a workflow for the classification revealed that the criteria for the categories reported in the paper were not exhaustive, likely because no projects in the original sample fell outside the reported categories. Unfortunately, it is not clear from the description in the paper how such additional projects ought to be categorized. Such ambiguity could be avoided by publishing an explicit analysis workflow.

Issues in adoption of eScience

The combination of growing large-scale public data sets and workflow tools such as Taverna and MyExperiment.org present a great opportunity for eScience on FLOSS and its development. We have already taken some steps in this direction with the creation of repositories of data and working papers, as well as the workflows described above. Our future work in this area includes building better interfaces to public datasets, creating metadata and ontologies for naming parts of datasets, (e.g., project and developer identifiers) and incorporating specifically social science data, such as content analytic schemas and annotated data (e.g., Heckman et al., 2007; von Krogh *et al.*, 2003), as well as records from interviewing and participant observation (subject to informed consent and appropriate human subjects review).

Of course, technical tools are only half of an infrastructure. To make the technology successful will require addressing a set of social issues that encourage or discourage its use. At the individual level, tool adoption requires developers to address issues such as ease of use and apparent utility, which we hope to demonstrate through study replications discussed above. For data, there are a broader set of questions. One set of issues involves policies for data curation to ensure that data (and analysis scripts) have the necessary documentation and are of acceptable quality to be reusable. In many cases, a considerable amount of tacit or domain knowledge is needed to make sense of the data, posing an obstacle to broader use. Such issues are particularly pressing for FLOSS data, which is mostly created as a by-product of work. A common objection to data sharing is concerns about the privacy of research subjects. The use of public data sources avoids some of these concerns, though FLOSS data poses interesting ethical questions about appropriate privacy policies for aggregated data that is already available elsewhere on the Internet. A related issue is the intellectual property implications of storing and redistributing such data.

Beyond these challenges, the most important issue will be developing motivations for individual researchers to participate, both in using and making available raw data, intermediate results and analyses. The effort began with ensuring easy access to data sets and demonstration of the utility of scientific workflow tools. Developers of the current repositories described above have made data openly available, driven in part by the prevailing ethos of openness in the communities they are studying, but for eScience approaches to be more broadly adopted, further incentives seem necessary. To understand these issues requires an institutional level of analysis, taking into account how current work practices are situated in a variety of settings and organizational structures. In an academic context, motivations for participation might include policies about rewards for sharing, e.g., citations, letters of recommendation or generalized reciprocity as well as more coercive enforcement via reviewing or funding policies. The onus also falls on editors and reviewers to shape the research community's practices. However, such efforts are difficult to organize in a multi-disciplinary field, such as FLOSS. Finally, the relationship between repositories and subjects themselves must be considered. For example, repositories might help FLOSS projects by playing an intermediary role between the projects and researchers, and even facilitate better access to data for project members as well as researchers.

Finally, a scientific infrastructure could have a significant impact by facilitating new collaborations (e.g., by creating virtual research centres). However, there are numerous challenges in doing so. At the individual level, people like to work independently or with a relatively small, collocated research group. Research that extends beyond one's own desktop or research group requires shared goals and direction for a larger research agenda which may be lacking in some contexts where the research community has not defined a set of accepted

grand challenges. The logistical challenges of coordination and distributed communication are important, but may ultimately prove secondary to the considerable mind shift required to be able to work effectively in larger-scale collaborative ventures.

Conclusion

In this abstract, we have described the nature of research on FLOSS development, as driven by the nature of FLOSS development itself. The research area seems ripe for the application of eScience tools, which will benefit the field by providing a shared base of data and tools and in the longer-run, more accumulation of results as researchers learn to build on each others' findings.

Acknowledgments

This research was partially supported by US NSF Grants 05-27457 and 07-08437.

References

Antoniades, I., Samoladas, I., Sowe, Sulayman K., Robles, G., Koch, S., Fraczek, K., et al. (2007). *Study of Available Tools* (EU Framework deliverable No. D1.1): FLOSSmetrics.

Christley, S., & Madey, G. (2007). Global and Temporal Analysis of Social Positions at SourceForge.net. In *Proceedings of the The Third International Conference on Open Source Systems (OSS 2007)*, Limerick, Ireland.

Conklin, M. (2004). Do the Rich Get Richer? The Impact of Power Laws on Open Source Development Projects. In *Proceedings of the Open Source 2004*, Portland, OR.

Conklin, M., Howison, J., & Crowston, K. (2005). Collaboration Using OSSmole: A repository of FLOSS data and analyses, *Symposium on Mining Software Repositories*. St. Louis.

Crowston, K., Howison, J., & Annabi, H. (2006). Information systems success in Free and Open Source Software development: Theory and measures. *Software Process—Improvement and Practice, 11*(2), 123–148.

English, R., & Schweik, C. M. (2007). Identifying success and tragedy of FLOSS commons: A preliminary classification of Sourceforge.net projects, *First International Workshop on Emerging Trends in FLOSS Research and Development*. Minneapolis, MN.

Heckman, R., Crowston, K., Eseryel, U. Y., Howison, J., Allen, E., & Li, Q. (2007, 11–15 June). Emergent decision-making practices In Free/Libre Open Source Software (FLOSS) development teams. In *Proceedings of the 3rd International Conference on Open Source Software*, Limerick, Ireland.

Howison, J., Conklin, M., & Crowston, K. (2006a). FLOSSmole: A collaborative repository for FLOSS research data and analyses. *International Journal of Information Technology and Web Engineering, 1*(3), 17–26.

Howison, J., Inoue, K., & Crowston, K. (2006b). Social dynamics of FLOSS team communications, *The Second International Conference on Open Source Systems*. Como, Italy.

Robles, G., Amor, J. J., & González-Barahona, J. M. (2005). Evolution and growth in large libre software projects. In *Proceedings of the The 8th International Workshop on Principles of Software Evolution*, Lisbon, Portugal.

von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in Open Source Software innovation: A case study. *Research Policy, 32*(7), 1217–1241.