Opportunities for eScience research on Free/libre Open Source Software

Kevin Crowston[1] and other workshop participants
Syracuse University School of Information Studies

Free/libre open source software (FLOSS) projects are an increasingly common and important approach to software development. In response, a growing number of researchers are examining this and related phenomena from a variety of perspectives. There is an opportunity for these researchers to mature into a research community supported by eScience tools. In this abstract, we present some conclusions of a US National Science Foundation-supported workshop on infrastructure for FLOSS research that describes a possible evolution of this infrastructure and community as a case study of the application of eScience ideas to a particular domain.

FLOSS is software developed and released under a license allowing inspection, modification and redistribution of the software's source[2], and often developed by voluntary communities of geographically distributed developers. Over the past ten years, FLOSS has moved from an academic curiosity to the mainstream, with a concurrent increase in the amount of research examining this phenomenon. Better support for this research is important for several reasons. First and foremost, FLOSS has become an important phenomenon to understand for its own sake. FLOSS is now an important social movement involving an estimated 800,000 programmers around the world (Vass, 2007) as well as a commercial phenomenon involving a myriad of software development firms, large and small, long-established and startup. On the user side, millions have grown to depend on FLOSS systems such as Linux not to mention the Internet, itself heavily dependent on FLOSS. In addition to its intrinsic merits, FLOSS has attracted great interest because it provides an accessible example of other phenomenon of growing interest. For example, many researchers have turned to FLOSS projects as examples of virtual work, as they are dynamic, self-organizing, distributed teams comprising professionals, users (von Hippel, 2001; von Hippel & von Krogh, 2002, 2003) and others working in loosely coupled teams.

To support FLOSS research and to help the diverse collection of researchers interested in FLOSS mature into a research community, we envision a shared infrastructure that will facilitate access to data, analyses, papers and other researchers. This infrastructure will include both a technological base as well as a set of enabling

[1] Hinds Hall 348, Syracuse, NY 13244 USA, +1 315 443–1676, crowston@syr.edu

[2] FLOSS software is usually available without charge (captured in a phrase commonly used in the community: "free as in free beer"). Much (though not all) of this software is also "free software", meaning that derivative works must be made available under the same unrestrictive license terms (captured as "free as in free speech", thus "libre"). We have chosen to use the acronym FLOSS rather than the more common OSS to acknowledge this dual meaning.

social mechanisms. Piece of this infrastructure already exist, providing an initial set of building blocks. For example, the field already has several repositories of raw data on FLOSS projects and coding, such as FLOSSMole (Howison *et al.*, 2006), the Notre Dame SourceForge repository (http://www.nd.edu/~oss/Data/data.html) and CVSanalY (Robles *et al.*, 2004). As well, the broader eScience community has seen developments in tools for eScience, such as the Taverna workflow system (http://taverna.sf.net/) and the MyExperiment social networking site (http://www.myexperiment.org/).

We envision a distributed technical infrastructure that will provide access to the various materials of science. A first component is raw data at various levels of analysis. FLOSS research is facilitated by the fact that much of the relevant data is "born digital", albeit as a by-product of work rather than as scientific data. For each project, we can provide access to raw data such as email archives, trackers, source code, documentation, release information, popularity, dependencies, download counts, as well as documentation of the project's history (e.g., releases, tool use, hosting). Data at the developer and firm level could also be shared. A particularly interesting kind of data is the original source code. To make all of these kinds of data usable though will require efforts to create better meta-data, such as the data dictionaries, provenance information and even just a catalog of what data are available. As noted above, several projects already collect some of this data, but there is still a lot to do, particular in the area of better meta-data.

In addition to raw data, an infrastructure should enable sharing research products such as research methods, ontologies and results. Researchers ought to be able to easily share information such as sampling frames and commonly studied samples of projects, and analysis workflows, both specific to a particular research study or as reuseable components, e.g., for data cleaning or sampling. At the project level, we need a census of FLOSS projects to establish the universe of study. Research results also should be shareable, either as new data sets or as annotations on existing data. For example, email data might be coded for various theoretical concepts and then shared as a starting point for further analyses. Source code might be analyzed for code structure, complexity or to illuminate project interdependencies.

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 the use of the infrastructure. One set of issues involves policies for data curation to ensure that data have the necessary documentation and are of acceptable quality to be resuable. Funding will have to be obtained to preserve and migrate data and to make it accessible to researchers. FLOSS data poses interesting ethical questions that must be addressed, such as appropriate privacy policies for data that is already available elsewhere on the Internet. A related issue is the intellectual property concerns about storing and redistributing such data. A key issue will be developing motivations for individual researchers to participate, both in using and making available data. Such motivations 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 conference policies. Finally, the relationship between repositories and the FLOSS projects themselves must be considered. Repositories might help FLOSS projects be being an intermediary between the projects and researchers and even facilitate better access to data for the projects as well as researchers.

In conclusion, FLOSS researchers seem well positioned to adopt eScience tools and practices and in so doing, evolve from a diverse group of individual scientists to an integrated research community with shared data, concepts, methods and tools. However, there is much work to be done in creating both the technical and social infrastructure for such sharing. We hope to learn from the successful efforts of others in implementing our vision for improving the field.

References

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

Robles, G., Koch, S., & González-Barahona, J. M. (2004). Remote analysis and measurement of libre software systems by means of the CVSanaly tool. In *Proceedings of the 2nd ICSE Workshop on Remote Analysis and Measurement of Software Systems (RAMSS), 26th International Conference on Software Engineering*, Edinburgh, Scotland.

Vass, B. (2007). Migrating to Open Source: Have No Fear, *3rd DoD Open Conference: Deployment of Open Technologies and Architectures within Military Systems*. Vienna, VA.

von Hippel, E. (2001). Innovation by user communities: Learning from open-source software. *Sloan Management Review*(Summer), 82–86.

von Hippel, E., & von Krogh, G. (2002). *Exploring the Open Source Software Phenomenon: Issues for Organization Science*. Cambridge, MA: Sloan School of Management, MIT.

von Hippel, E., & von Krogh, G. (2003). Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science. *Organization Science, 14*(2), 209–213.