Towards a data and workflow collaboratory for research on free and open source software and its development
A poster submission for iConference 2008

(Author details redacted)

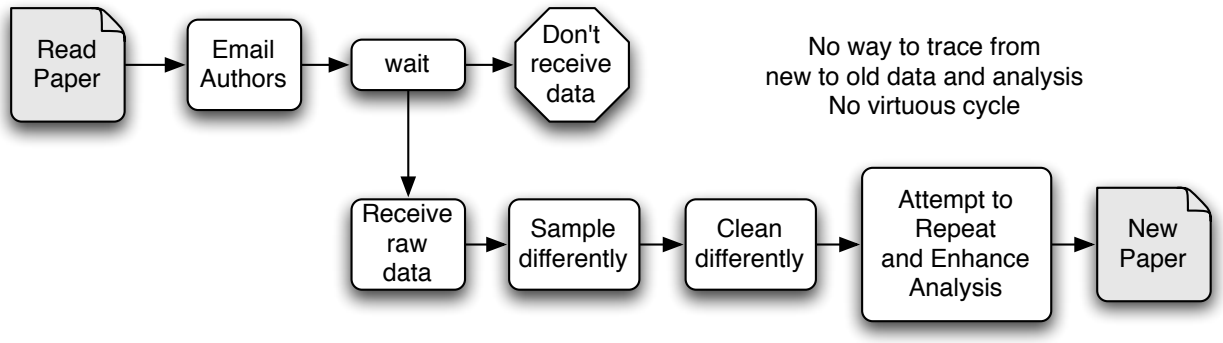
29 October 2007

FLOSSmole is a collaborative data repository which collects and provides data for research on Free/Libre Open Source Software (FLOSS) and its development by online, distributed teams. The data is used by a research community that studies diverse questions from the evolution of software to how these groups make decisions, use various media and manage change over time (Scacchi, 2007). This multi-disciplinary research community includes researchers both inside and outside iSchools, from many disciplines including software engineering, organizational studies, information systems and sociology, as well as corporate researchers.

Since May 2007, we[1] have been working to development this repository into an e-Social Science infrastructure capable of supporting, storing and publishing not only data but analyses organized into scientific workflows, as envisioned in the NSF reports on Cyberinfrastructure (Atkins, 2003; NSF Cyberinfrastructure Council, 2007). Further we are developing a pre-print repository which will enable bidirectional links between the data, workflows and published papers. The goal of the project is to facilitate collaboration to improve the reproducibility and consistency of research, collaboratively building a body of cumulative knowledge (Borgman, 2007; Finholt and Olson, 1997).

[1]References to our FLOSS research have been excluded for review

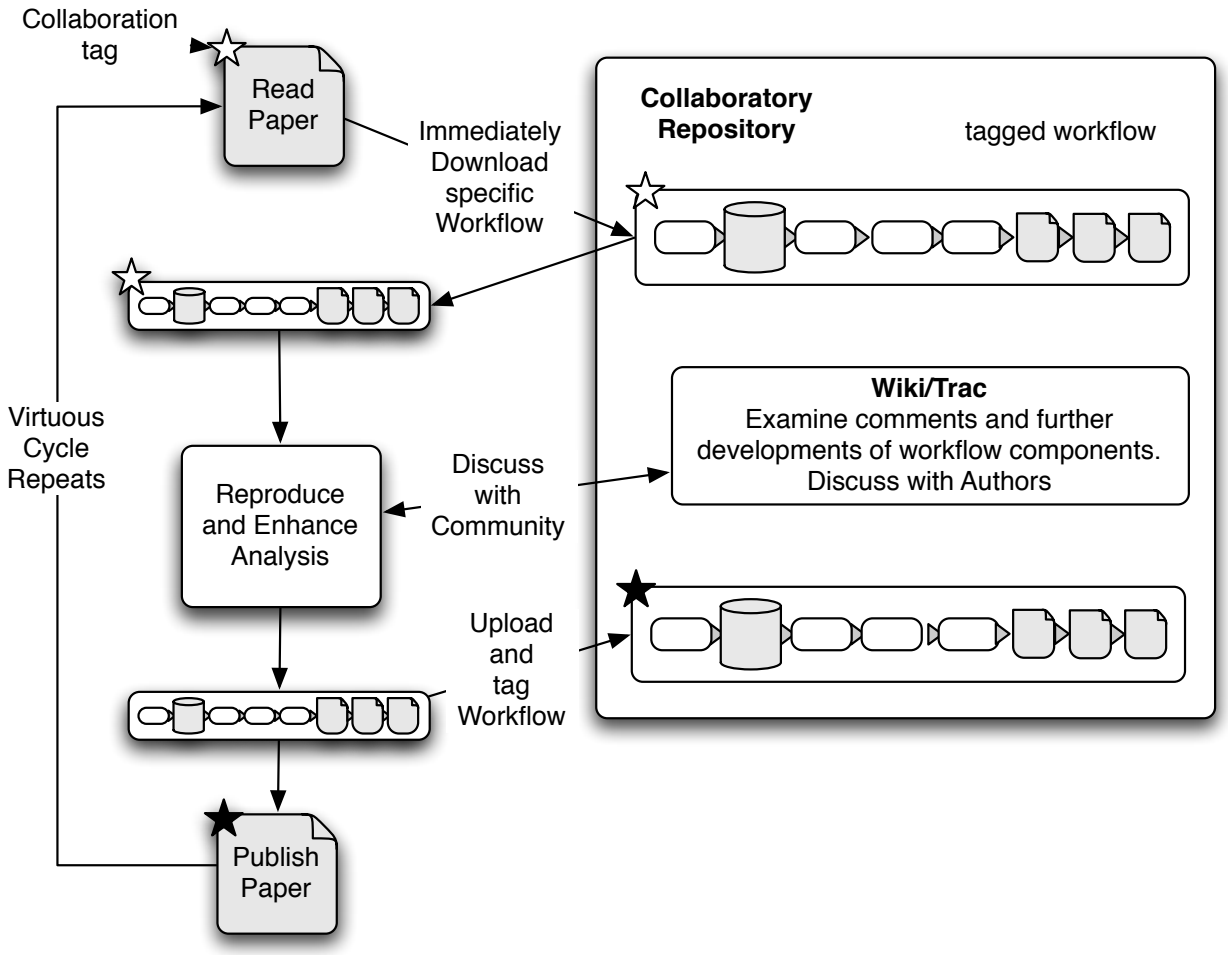
Figure 1: *Current research practice has significant difficulties*



The current social science research on FLOSS has relied on several different kinds of scientific evidence, including the archives created by the FLOSS developers, versioned code repositories, mailing list messages and bug and issue tracking repositories. 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 data are by-products, they are not in a form that is useful for researchers. Instead potentially useful data is locked up in HTML pages, CVS log files, or text-only mailing list archives. Furthermore because FLOSS projects are hosted in a variety of "forges" (of which Sourceforge is but the largest of dozens) or individual websites, these problems are multiplied. Finally different projects use similar services in quite different ways, leading to inappropriate comparisons. Research projects expend significant energy collecting and re-structuring these archives for their research. This process is repetitive, wasteful and produces inconsistent results.

The situation with analyses and workflows is even more inconsistent and repetitive, with different research groups performing similar sampling, data cleaning and analyses steps but unable to share or to build on each others' work. Our project is learning from e-Science infrastructure in the natural sciences, which provide shared databases for data such as genome sequences and astronomical observations, often hosted under the purview of the NSF's Ter-

Figure 2: *Virtuous cycles enabled by research workflow sharing*



raGrid project. They have developed graphical desktop software, such as Taverna[2] and Kepler[3], to link together these data with analytical transformations which can be made available as downloaded components or web-service gateways to distributed computational resources ('grid computing'). Increasingly it is also possible to include commercial services, such as Yahoo's Pipes or Amazon's storage and computational resources.

Crucially the workflows can be stored and shared (as XML files) and an online service is provided to 'stamp' a particular workflow configuration with a unique identifier for later

[2]http://taverna.sourceforge.net/

[3]http://kepler-project.org/

reference. Each 'run' of the workflow produces a record of the transformations undertaken, preserving the provenance of the analytical results. It is hoped that our project will facilitate the use and sharing of such workflows, allowing FLOSS researchers to re-use appropriate and tested components, freeing them to concentrate their efforts on novel contributions ('return from scripting to science'). Such documentation and re-use is valuable both between research groups but also within individual research groups over time, as collaborators and students come and go.

As useful as these natural science systems are, their application to social science requires significant adaptation (Berman and Brady, 2005), which our poster presentation will highlight for the benefit of the multi-disciplinary iSchool audience. For example, the FLOSS research community commonly draws on content analytic techniques to investigate textual archives, such as mailing lists. This content analysis can be both positivist (abstracting observations, often by producing counts or sequences of observed concepts) or interpretative (focused on accessing the meaning for participants). These 'marked-up' texts, as well as the content analytic schemes developed, are a resource which could be valuable to other researchers. However it is vital that the 'workflow' of their production is also clearly described together with intermediate artifacts like inter-coder reliability statistics. Our project is exploring appropriate ontologies, like those of the Data Documentation Initiative[4], and will be able to present preliminary results in February.

Finally, publications refer to data and measures using inexact and conflicting descriptions, complicated by the fact that researchers often draw on the same raw data to operationalize different concepts. It is hoped that our research will create a controlled namespace, allowing researchers to move both from publications to specific data and workflows, and to see quickly how particular data or analysis techniques have been used in published papers.

In order to demonstrate the value of such techniques, and to kick-start their use in this research community, our project will engage in digital curation, selecting six important papers

[4]http://www.ddialliance.org/

which will be replicated using the adapted e-Science techniques. The data selections and analysis components so created will be made available as base infrastructure and introduced at an NSF-funded FLOSS research community workshop to be held just prior to the iSchool conference.

In summary, the proposed poster will stimulate discussion about multi-disciplinary scientific ecosystems, requirements for adapting Cyberinfrastructure and e-Science techniques for the social sciences and provide a template for those in the iSchools community seeking to enhance collaboration and quality in their research communities.

Bibliography

Atkins, D. (2003). Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.

Berman, F. and Brady, H. (2005). Final report: NSF SBE-CISE workshop on cyberinfrastructure and the social sciences. Available at www.sdsc.edu/sbe/.

Borgman, C. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press.

Finholt, T. A. and Olson, G. M. (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science*, 8(1):28–36.

NSF Cyberinfrastructure Council (2007). Cyberinfrastructure vision for 21st century discovery. NSF Report 0728.

Scacchi, W. (2007). Free and open source software development: Recent research results and methods. In Zelkowitz, M., editor, *Advances in Computers*, volume 69. Elsevier Press.