Toward a Framework for Evaluating Software Success: A Proposed First Step

Stan Ahalt (ahalt@renci.org), Bruce Berriman, Maxine Brown, Jeffrey Carver, Neil Chue Hong, Allison Fish, Ray Idaszak, Greg Newman, Dhabaleswar Panda, Abani Patra, Elbridge Gerry Puckett, Chris Roland, Douglas Thain, Selcuk Uluagac, Bo Zhang.

Software is a particularly critical technology in many computational science and engineering (CSE) sectors. Consequently, software is increasingly becoming an important component in the evaluation of competitive grants and the execution of research projects. As a result, software can be viewed as a scholarly contribution and has been proposed as a new factor to consider in tenure and promotion processes. However, existing metrics for evaluating the capability, use, reusability, or success of software are sorely lacking. This lack of software metrics permits the development of software based on poor development practices, which in turn allows poorly written software to “fly under the radar” in the scientific community and persist undetected. The absence of evaluation by knowledgeable peers often leads to the establishment and adoption of tools based on aggressive promotion by developers, ease-of-use, and other peripheral factors, hindering the sustainability, usefulness, and uptake of software and even leading to unreliable scientific findings. All of these factors mean that addressing the current lack of software evaluation metrics and methods is not just a question of increasing scientific productivity, but also a matter of preventing poor science.

As a first step toward creating a methodology and framework for developing and evolving software success metrics for the CSE community, we propose the creation of a software “peer-review group.” This group, comprised of grant recipients funded to develop sustainable software, would meet periodically to evaluate their own and each others’ software, developing and refining success metrics along the way. We envision the group as a pilot test for a potential larger-scale effort to establish a more formal framework for software success metrics and evaluation.

Framing Success Metrics
Our perspective on framing software success metrics arose from a breakout session held at a recent NSF-funded workshop attended by more than 75 Software Infrastructure for Sustained Innovation (SI2) principal investigators. The breakout team identified the need to create a methodology and framework for academic software success metrics, brainstormed factors to consider in developing such a framework, and outlined the actionable steps needed to advance this effort. The idea of a software review group was introduced in these discussions, and possible outcomes—presented briefly here—were discussed. We believe further discussion by Computational Science and Engineering Software Sustainability and Productivity Challenges (CSESSP) workshop attendees will help to further develop these ideas and emphasize the importance of framing software success metrics as an integral part of developing a sustainable software ecosystem.

The Need to Evaluate Software Success
On the whole, the development of research software in academia, government, and national labs trails the rigor of industry-developed software. Incentives and measurements of what constitutes successful software differ among and within these sectors, yet all are ultimately part of the same software ecosystem. Generally speaking, successful software must be reliable, sustainable, have value to the target user community and beyond, and provide outcomes that are meaningful to societal stakeholders. Sound software development and engineering practices lead to sustainable software. Stakeholder adoption, use, and reuse of software create feedback loops that further enhance software success. To improve the productivity and sustainability of research software and the research communities it supports, we should be able to objectively measure what makes software successful—or not.
Factors to Consider

There are multiple dimensions to consider in developing an effective methodology and framework for evaluating software success. One dimension relates to the factors that contribute to software success, such as criticality, usability, performance, functionality, availability, and scientific impact. These terms may have different meanings in different fields; for example, usability may mean something different for networking software than it does for security software. Another dimension relates to the types of outcomes we might want to measure, such as the value of the scientific contributions of a grant or project, the value of the products of a grant or project (i.e., the value of the software), or the nature of the team’s “community conduct” (e.g., its value to the software ecosystem). Another relates to defining needs: for example, what is it that funders, researchers, or the broader community need to know in order to inform better decisions and improve sustainability? Finally, we must develop robust metrics to address these dimensions, inform project goals, and empower software creators, researchers, funders, and others to effectively evaluate software.

Next Steps

To begin to develop and evolve a software evaluation framework, we propose establishing a peer review group—an organization of representative stakeholders who will self-review software works created by their respective communities. This group would effectively constitute a pilot program to inform the feasibility, scope, and approach of a future, larger effort to establish and refine a framework for sustainable software metrics. At a minimum, this group would give its members an opportunity for regular review and enhance their own self-improvement processes. If successful more broadly, the group would help to characterize key challenges in software evaluation, define and refine evaluation criteria, and lead to a more informed approach to software development and evaluation for the CSE community as a whole.

Recently, this idea was discussed at the WSSSPE workshop.\textsuperscript{iii} We believe further discussion of this idea at the CESSSP workshop would refine and inform our approach and help to generate momentum toward achieving better software evaluation approaches. Examples of questions that warrant further exploration include:

- How will peer-review of code and corresponding peer-review of software success metrics be performed and tied together across different research domains?
  - What type of infrastructure will best facilitate this?
- How will ever-increasing open participation by the community be nurtured and measured?
- What governance structure will be put in place to ensure shared community governance of the software and metric peer-review process?
- How will factors of software success be evaluated: e.g. criticality, value, sustainability, usability, reusability, performance, functionality, capability, availability, scientific impact, usefulness, reliability?
- Can software success metrics be fit into a common template?
- How will the process for submitting software for evaluation be managed?
  - Should all group members be required to regularly submit their own software?
- What types of documentation or outcomes would be useful toward developing a larger-scale metrics framework?

\textsuperscript{1} Stan Ahalt, RENCI, University of North Carolina at Chapel Hill; NCDS, ahalt@renci.org
Bruce Berriman, California Institute of Technology, gbb@ipac.caltech.edu
Maxine Brown, University of Illinois at Chicago, maxine@uic.edu
Jeffrey Carver, University of Alabama, carver@cs.ua.edu
Neil Chue Hong, Software Sustainability Institute, University of Edinburgh, N.ChueHong@software.ac.uk
Allison Fish, University of California, Davis, aefish@ucdavis.edu
Ray Idaszak, RENCI, University of North Carolina at Chapel Hill, rayi@renci.org
Greg Newman, Colorado State University, Gregory.Newman@ColoState.Edu
Dhabaleswar Panda, The Ohio State University, panda@cse.ohio-state.edu
Abani K. Patra, University at Buffalo, abani@buffalo.edu
Elbridge Gerry Puckett, University of California, Davis, egpuckett@ucdavis.edu
Christopher Roland, North Carolina State University, cmroland@ncsu.edu
Douglas Thain, University of Notre Dame, dthain@nd.edu
Selcuk Uluagac, Florida International University, suluagac@fiu.edu
Bo Zhang, Indiana University, zhang416@indiana.edu

Report from the National Science Foundation-funded workshop held February 17-18, 2015, at the Westin Arlington Gateway in Arlington, Virginia for Software Infrastructure for Sustained Innovation (SI2) Principal Investigators, [http://dl.acm.org/citation.cfm?id=2764957](http://dl.acm.org/citation.cfm?id=2764957).