FOSSProF Final Report Template

Project Overview

- SteamRoller: replicable and scalable empirical humanities
- Project Summary: Improvements to the SteamRoller framework to operate smoothly on a wider variety of grid engines, including JHU's new DSAI infrastructure, to refactor and package the library, and to make several augmentations to functionality.
- Target Audience: Applied computational research, particularly in data-driven machine learning.
- Code Repository: https://github.com/comp-int-hum/steamroller

Project Activities and Progress

- Work Completed: SteamRoller is now fully functional on the open-source grid system used by JHU (and many others), has been standardized to follow best practices for Python packaging, and includes useful on-ramps for new users to initialize projects. Several downstream users are using it actively and providing feedback while also publishing research outcomes whose associated experimental repositories are SteamRoller-based. Several hooks for future work have been implemented in the code base, in particular by intercepting the Scons invocation so it can be customized directly.
- Technical Milestones: New users include undergrad and grad students at JHU in the humanities and engineering, and researchers from Cambridge, Cornell, and Hamburg. A new major version (3.0) was released after the code refactoring. An afternoon tutorial was run in the early Summer and will be developed into a full course for Spring 2026.
- Challenges and Solutions: A simple time crunch was the major hurdle, due to unexpected responsibilities for the PI (personal and DSAI-related), but the surface area of the project meant it was possible to find multiple paths to research impact.

Outcomes and Impact

- Project Impact: The initial proposal emphasized the research impact SteamRoller has had over the past several years just as a fairly casual project, providing the backbone of over a half-dozen published outcomes. The best measure of the grant's impact is that over the past ~9 months, that number has effectively doubled, and the geographic diversity of users expanded beyond JHU.
- Community Engagement: Aside from on-boarding new users directly, the main community engagement has been at conferences focused on research outcomes driven by SteamRoller, such as TADA2023, EACL2024, ACL2024, and the upcoming EMNLP2024. These are the top-tier venues for NLP and language-centric machine learning.
- Sustainability / Future Plans: SteamRoller will be robustly supported as long as researchers are using it, and that community has grown substantially over the past year. Moreover, impact will continue to dramatically increase now that it operates more cleanly on JHU's key infrastructure, and pedagogical efforts are under way.

- Lessons Learned: The primary takeaway is the need to prepare robust introductory materials once SteamRoller is ready for general use at the advanced undergraduate level. On-boarding people directly is feasible if they have a moderate intuition for the concepts, but a substantial population would benefit from a standardized overview and hands-on walkthroughs.
- Attachments: In addition to repository link above, see all 2024 publications:
    - https://scholar.google.com/citations?hl=en&user=9oyfC6UAAAAJ&view_op=list_works&sortby=pubdate
    - https://scholar.google.com/citations?hl=en&user=bMk-rOQAAAAJ&view_op=list_works&sortby=pubdate
    - https://scholar.google.com/citations?hl=en&user=8JDH5b0AAAAJ&view_op=list_works&sortby=pubdate