Delivery Mechanism and Community Usage Metrics

Elements: Improving tools based on data-description standards for gigabyte-scale data sets

PI: Amy Roberts

The long-term goal of this work is to increase the accessibility of science analysis for both active researchers and science students. There are three primary products associated with this work:

1. Community data-access software that allows scientists to analyze binary data in any format.
2. Documentation that includes installation instructions, example analyses, and learning materials that cover the basic computing skills needed to use the software.
3. Workshops that bring together scientists and developers.

In Year 1 the PI expects to meet the following metrics of community involvement:

- Between 15 and 20 scientists register for the Data Access workshop. These scientists are expected to predominantly come from the nuclear physics community. The PI expects all attendees to register after specific invitation.
- Between 5 and 10 scientists register projects with the Open Science Framework or similar platform to work on analysis of their data collaboratively
- An initial release of the improved software and the XIA library is made. Both have basic testing coverage, are tested automatically upon commit, and have initial guidelines for contribution.
- An initial release of the skills documentation has been made. Inexperienced students who try to follow the analysis tutorial are able to find answers to some of their questions but most are expected to be unable to complete the tutorial without expert assistance
- A roadmap for the data-access library is published post the workshop

In Year 2 the PI expects to meet similar goals, but with more community support. The PI expects one or two attendees to find out and register for the conference from someone outside the group and for most to register after specific invitation. The PI expects more researchers to pre-register their analysis on the Open Science Framework. It is expected that the releases of the software and documentation will improve the community-identified pain points significantly, and that there will be continued involvement in roadmap discussions during the workshop.

In Year 3 the PI expects to see broader community support. Most notably, it is expected that the data-access software will be cited in 2 to 5 peer-reviewed papers that present scientific results.

The PI expects that issues with the software and documentation will be relatively minor and that the ecosystem, now well-developed, will promote approximately half the scientists to pre-register their analysis before the workshop. The PI expects that the roadmap discussion will involve multiple viable plans for the sustainability and continued maintenance of the software.

The PI feels that permissive open-source licenses align best with the goal of broad adoption. Because this work will only be sustainable and impactful with the broadest possible community adoption, permissive open-source licenses will be preferred wherever possible. The PI does not anticipate spending NSF resources on closed-source software.

Deliverable	Mechanism	Metric
Improve existing data-access tools based on the Katai Struct description language to be a viable tool for gigabyte-scale data analysis		
Add Katai Struct target for awkward-array	Github, Gitlab, Zenodo	citations of software by science-result papers
Build a library for XIA data based on the common-access code	Github, Gitlab, Zenodo	citations of software by science-result papers
Prototype an analysis library for SuperCDMS based on the common-access code	CDMS will consider releasing this code if it becomes the basis for ongoing analysis	improvements to the common-access code and associated libraries such as awkward-array based on analysis experience
Build an active community of users and developers for standards-based science software		
Hold a yearly workshop for community learning and development	Meeting materials and selected recordings (??) will be archived on the Open Science Framework	Increasing numbers of attendees who are not specifically solicited by the group
Publish materials with clear and complete instructions for contributing to the project	released as part of the code base on github, gitlab, Zenodo	increase in who contributes commits to the code
Build testing into all released code to allow easier contributions from new developers	released as part of the code base on github, gitlab, Zenodo	increase in who contributes commits to the code
Maintain an issues forum and a discussion forum for the project	issues forum available on gitlab for each project; Discourse provides forum hosting for open-source projects	an increase in questions directed at my team and an increase in discussion between new participants

Table 1. Reqeusted salary support from all received and pending NSF grant applications, in months. Note that "year 1" in this text refers to the suggested start date of this proposal, 11/1/2019.