Proposal Review 4 : 1931382

Back to Proposal

Agency Name:	National Science Foundation
Agency Tracking Number:	**1931382**
Organization:	
NSF Program:	Software Institutes
PI/PD:	Roberts, Amy
Application Title:	Elements: Improving tools based on data-description standards for gigabyte-scale data sets
Rating:	Fair

Review

Summary

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to intellectual merit.

Short Description
This research proposes to enhance existing data analysis software that is limited in the nuclear physics community. Data analysis tools need to handle gigabyte scale data that is common in this field. Custom format data is an input to this software, and although existing packages support this, they have poor performance. The PI proposes to enhance gigabyte data performance through a Python API, building a community though yearly workshops, and increasing access to nuclear physics research to undergraduate students.

In the context of the five review elements, please evaluate the strengths and weaknesses of the proposal with respect to intellectual merit.

Strengths
ò The tools developed will help the community in research related to super cryogenic dark matter search, symmetry testing, origin of elements in the universe, and radiation monitoring
ò Tools will enhance access to binary format data through a user specified format via a data description language. This will be fully documented
ò The work will be integrated incrementally year over year through workshops, where each year at the end of the workshop, a new roadmap can be created for the following year

Weaknesses
ò The user stories provided by the PI are typical in all fields of science, not just nuclear physics, and this does not seem to motivate the work very well. Are these fictional stories? The reviewer will assume that is the case, since there are not references
ò Many of the goals and the proposed work keeps repeating itself in all sections of the paper
ò Although the PI will build on infrastructure that is already available, the proposal is focused on developing documentation, training and workshops rather than the development of the software itself
ò The technical aspects of the proposal are significantly lacking. In the case where the PI discussed the potential impacts of this work, although it is understandable that the work should begin by leveraging existing collaborations with XIA and CDMS, the focus should really be on developing the generic data description language first
ò There is no justification as to how developing the analysis tool that is compatible with any data format can be carried out within one year, there just isn't enough evidence to support this in the proposal. Where are the architectural diagrams, etc.
ò The end-to-end testing described by the PI is not supported. To my best understanding, only students will be doing user testing (as evidenced by the lifecycle in Figure 1), but this is certainly not end-to-end. Where are the plans to do unit, interface, component, acceptance testing?
ò The list of initial requirements provided by the PI do not focus on the functional and non-functional aspects of the software. This is especially needed is you do at a minimum acceptance testing. The PI needs to pair up functionalities to test cases

In the context of the five review elements, please
evaluate the strengths and weaknesses of the proposal with respect to broader impacts.

In the context of the five review elements, please evaluate the strengths and weaknesses of the proposal with respect to broader impacts

Strengths
ò The proposal to support any user described data format for gigabyte sized information is potentially beneficial assuming this capacity is not currently supported by other software
ò Software is Open Access which allows for accessibility and contributions of a greater community

Weaknesses
ò Although better documented accessibility to software, especially for individuals in institutions that lack resources is a great goal, this is not enough. First and foremost, the software itself needs to be well architected and validated with possibly the help of a software engineer
ò There is no mention of broadening participation of minorities, diversity, gender diversification, etc.

Please evaluate the strengths and
weaknesses of the proposal with respect to any additional solicitation-specific review criteria, if
applicable

Please evaluate the strengths and weaknesses of the proposal with respect to any additional solicitation-specific review criteria:

Science-driven:
Strengths
Weaknesses
ò Although commendable to try to integrate undergraduate students into research by creating an easier set of software tools for their use, this is not enough to state this as a science driven approach. For instance, in this section of the proposal the PI states that she would provide a common toolset for analyzing data in any format, so I would focus on why this is important and specifically how this will be done

Innovation:
Strengths
Weaknesses
ò Leveraging existing technology is a smart way to increase the chances of better software, but this in itself is not innovation. The PI should focus on a description of the proposed data description language for any format of data. This is where the innovation is

Close collaborations among stakeholders:
Strengths
Weaknesses
ò It is hard to see how the PI can motivate existing cyberinfrastructure experts to use this software. I understand you can start with close collaborators, but be specific as to how the software can attract mature users, not just students

Building on existing, recognized capabilities:
Strengths
ò There is evidence to suggest that JSON and XML produces additional secondary files that make their size bigger than binary files. This would make

these technologies unusable or impractical for gigabyte scale data. It would be nice to provide an example or a reference to the sizes being discussed here. For example, there are techniques such as memory mapped files that alleviate this problem. However this recognition by the PI is aligned with using the Katai Struct software

ò Using the Python ecosystem is good strategy as it is widely used and understood

Weaknesses

ò It is not clear how the PI will build on the Katai Struct compiler software. There is no evidence in the staff working on this proposal that the necessary skillsets are there. Additional information is required to learn about how the awkward-array data structure will be redefined. More information is needed here

Project plans, and system and process architecture:

Strengths

ò The development od documentation, quick resolution to questions, ease of installation instructions, links to additional resources, instructions for developing and testing, etc. are great goals for this software and, in the opinion of the reviewer, very attainable

ò Delivering software that can be maintained by the community

ò The software tools and processes described (in page 10) are well thought out and "fit" this type of project well. The PI proposes an Agile approach with incremental prototyping.

ò The reproducibility section describes a first step to producing reliable and replicable results.

Weaknesses

ò Table 1 and Figure 1 do not show evidence of tasks for end-to-end testing. This should be significant in this project that is focused on improving usage

ò The PI describes a functional requirement to write a "size" function for the Katai Struct API, then writing a "compiler" that translates concepts (in Scala) to a target language and a runner to test said language requires more than a sentence. This is an opportunity to describe how this solution improves on the state of the art, but there are no diagrams, performance goals, potential implementation details, etc. Modifying a compiler is not a simple task

Deliverables:

Strengths

ò An improved data access library that is scalable to gigabyte scale data

ò Lots of documentation, tutorials, etc.

Weaknesses

ò Deliverables are mostly documentation, workshops and training materials. To make the software tools extensible and maintainable, the PI should also consider delivering architectural diagrams, information on testing strategies, etc.

Metrics:

Strengths

ò Tracking the number of developer contributions to the code base is probably more important to the number of citations. I would focus on the former. Citations will only come with adoption of this software. Hopefully, the contributions come from scientists that are also outside the XIA community

Weaknesses

ò It would be nice to see some QA metrics in terms of LOC, bugs per 1K LOC, branch coverage, static analysis of code smells. There are important (of the shelf) tools that can be run. It is especially important to community released tools

Sustained and sustainable impacts:

Strengths

ò The most important aspect in this section, is the potential reduction in time spent by developers when dealing with data access software.

Alignment with Directorate Specific Priorities:

The goals of the "elements" program targets small services that advance some parts of science and engineering. I believe the PIs is more focus on the training aspects than the technical.

Summary Statement

This research discusses an improvement to dealing with gigabyte scale data. The proposal seems more focused on the documentation, training and workshops rather than developing the technical aspects of the proposal. The reviewer would have liked to see an emphasis on the technical contributions such as the configurable data formatting, the improvements to data structures, the development of compilers, etc. that truly show an improvement over existing tools and software. This is missing. It is also unclear that the staff and stated salaries match with the proposed work. For example, wages of $18 per hour seem unrealistic for graduate students. What about tuition waivers, health, etc?