Delivery Mechanism and Community Usage Metrics

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

PI: Amy Roberts

Deliverables

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 deliverables associated with this work:

1. Foster the awareness and adoption of existing data-description languages
2. Improve existing data-access tools based on the Katai Struct description language to be a viable tool for gigabyte-scale data analysis
3. Build an active community of users and developers for standards-based science software

Year 1

- A whitepaper is published describing data description languages and giving use-cases
- The Open Science Framework and Figshare are contacted. We find out if they're interested and if so, get a contact person.
- 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 improvements to the core library and plans for additional, useful libraries is published and updated post the workshop

Year 2

- Between 15 and 20 scientists register for the Data Access workshop. Some diversity of discipline is expected, although most are expected to predominantly come from the nuclear physics community. 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.

- 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 highest-priority helper code is made. It has basic testing coverage, are tested automatically upon commit, and have initial guidelines for contribution.
- The highest-priority improvements for the core library are released. The contribution guidelines and instructions are well-tested. There has been at least one request for an improvement or feature from the community that has been either fixed by my team or another contributor.
- The most common failings of the skills documentation have been addressed. Inexperienced students who try to follow the analysis tutorial get stuck on these issues less frequently. Most are expected to be unable to complete the tutorial without expert assistance
- The roadmap discussion has more participants than in year 1

Year 3

- Between 15 and 25 scientists register for the Data Access workshop. Some diversity of discipline is expected, although most are expected to predominantly come from the nuclear physics community. The PI expects at least three attendees to find out and register for the conference from someone outside the group and for most 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
- The highest-priority helper code is close to stable. The contribution guidelines and instructions are well-tested. There has been at least one request for an improvement or feature from the community that has been either fixed by my team or another contributor.
- The highest-priority improvements identified on the road map for the core library are released. The contribution guidelines and instructions are well-tested. There has been at least one request for an improvement or feature from the community that has been either fixed by my team or another contributor.
- The basic-skills documentation has resources or recommends resources that address most of the questions that arise when inexperienced students try to follow the analysis tutorial. Most students need some help but all are able to make a plot based on data
- The roadmap discussion has more participants than in year 2

Deliverable	Mechanism	Metric
Foster the awareness and adoption of existing data-description languages		
A paper describing data-description standards that gives use cases for Kaitai Struct and DFDL	Open Access Journal, Conference presentation	Citations
Development of community standards	Research Data Alliance	Creation of a working group, whitepaper
Work with scientific data-sharing platforms and make them aware of data-description standards	Figshare, Open Science Framework	Support materials added to help scientists add description files to their projects
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
Create or improve existing tools that use the Katai Struct format for use in scientific analysis	Github, Gitlab, Zenodo	increasing downloads
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.

The proposed work will result in several software and documentation products, specifically

- Software based on kaitai-struct that interfaces with the data structures developed by IRIS-HEP that are optimized for science data analysis. This includes end-user and developer documentation.
- An additional release of the XIA python-based analysis library that uses the above code for easier analysis of large data sets. This includes end-user and developer documentation.
- Releases of "helper" libraries that are useful to the scientific community, such as the Kaitai Struct visualizer that reads and nicely displays binary data but needs improvements to work well with gigabyte-scale files. This includes end-user and developer documentation.
- A project website that summarizes and links to the above software and that provides educational resources for fundamental concepts in scientific computing.

Small, example data files may be prepared for inclusion with the software as test cases.

Metrics

Small data files meant to allow testing of the software will be stored in their original, custom binary formats. The descriptions of these formats will be stored in ASCII text files following the Katai Struct data description standard.

The point of the software developed under this work is to provide easy access to data stored in non-standard formats; documentation for the use of this software will be heavily tested.

Access to Data and Data Sharing Practices and Policies

The software created by this project will be publicly available for download from a cloud-based repository host such as github or gitlab. Additionally, all software products will be registered, archived, and available for download on Zenodo.

Software products will be publicly available throughout their development; releases will be used to guide users to stable versions. All releases of the software will all be archived and available on Zenodo.

Papers related to the software products will generally be preceded by a software release; the availability of the software products is otherwise independent from publications.

Individuals and organizations who request the software will be directed to download the code through the public channels.

Permissive open-source licenses (MIT, Apache, CC-BY 4.0) allow others to re-use the software but does not require that they grant the same license to users of their product. This makes it easier for companies to use the software since they're not required to share the source code.

Copyleft open-source licenses (GPL, BSD) allow others to re-use the software but requires that they make the software available under the same or similar license terms. Copyleft licenses prioritize keeping source code freely available.

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.

The goal of the proposed work is to increase the accessibility of science analysis to the entire community. Therefore all products will be licensed to allow easy re-use for both non-commercial and commercial purposes:

- All written content on the project website and documentation and any images will be licensed under CC-BY 4.0.
- All code will be licensed with an open-source license that allows re-use of the code for commercial purposes.
- Articles written about the software use and development will be published as open access wherever possible; in every case preprints will be published either on the arXiv, figshare, and/or the Open Science Framework.