

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

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

Table 1. Requested 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.

The proposed work will not generate new data. Instead, this work will use already-collected data. Small, example data files may be prepared for inclusion with the software as test cases.

Metrics

The software source code and documentation will be stored in ASCII text files.

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 Kaitai 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.

Policies for Re-Use, Re-Distribution

Scientists who use the code produced as a result of this work will be asked to cite the version they use using the appropriate Zenodo DOI. All software will include citation instructions in the top-level README file.

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.

Archiving of Data

All software and documentation will be archived on Zenodo. Zenodo is a collaboration between CERN and OpenAIRE and has an operation plan for the next twenty years.

Zenodo saves all data to two physically distinct disk servers. For low-use data, they reserve the right to store the data to tape.