Data Management Plan

Expected Data

The proposed work will not generate new primary data. This work will focus on the utilization of already-collected data to prototype a data-delivery service. Small, example data files may be prepared for inclusion with the software as test cases. In addition, during the development process, ephemera such as code review comments, mailing list discussions, design documents, and other data in the form of communication materials will be produced.

The principle data collected during this project will have no privacy implications. Metadata for produced source code will be in the form of git changesets and code review comments, both of which are industry standard. Licensing of source code, as detailed in the proposal, will adhere as much as possible to "permissive," non-copyleft BSD licenses, which allow re-use, re-distribution and the production of derivatives. There are a handful of exceptions to this, most notably in the contributions that would be made "upstream" to Kaitai Struct, which will be licensed under the terms of the upstream license (GNU General Public License, or GPL) which is an open source, "copyleft" license.

We anticipate the production and distribution of sample data. Data developed for this purpose will either be generated and licensed under a Creative Commons Zero license or we will solicit access to it from community members and distribute it only under their terms.

Design documents developed in the course of this proposal (including YTEPs, as described in the proposal) will be versioned and available freely in both original source form and in rendered form on the project website. Summary documents describing the technical process and outcomes of the project (so-called "code papers") will be published through an open access venue such as the Journal of Open Source Software.

The primary data products of this work will be:

- 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, source code and testing artifacts.
- Updates to the yt analysis package; this will include tutorial materials, source code, and design documents.
- Documentation for standing up and connecting the data-delivery service to a generic experiment.
- Example analyses that use the data-delivery service.
- A project website that summarizes and links to the above software.

Data Format

The software source code and documentation will be stored in ASCII text files, versioned on GitHub, with additional off-site backups. We do not anticipate utilizing either a contributor license agreement (CLA) or a developer certificate of origin (DCO) but we will explore their necessity with the relevant technology transfer officers at our local universities.

Small data files meant to allow testing of the software will be stored in their original, custom binary formats; this is by-design, as the purpose of the project is to facilitate access to the unmodified, original data formats. The descriptions of these formats will be stored in ASCII-encoded text files following the Katai Struct data description standard. These will be versioned and documented in a central GitHub data format repository. At regular release intervals, we will "tag" the current state of these and publish them to an external repository such as Zenodo.

The point of the software developed under this work is to provide easy access to data stored in non-standard formats; our documentation will be versioned and utilized in our testing infrastructure.

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 private industry to use the software, as redistribution does not require preservation of source availability.

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.

Permissive open-source licenses align best with the goal of broad adoption. Because this work will only be sustainable 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.