EIC Detector-1 Software Decision

Topic: Data and Analysis Preservation

Point of Contact: Kolja Kauder (kkauder@bnl.gov)

Discussion Date(s): July 13, 2022

Meeting Link: https://indico.bnl.gov/event/15646/

Overview:

Data & Analysis Preservation (DAP) traditionally aims at **reproducibility**, at continuing on-going analyses, and at outreach and education. Starting about two decades ago, its role has expanded to include **reusability** and even **reinterpretation** of existing data and analyses.

The EIC and its detector(s) are in the rare, if not unique, situation of having committed to DAP already in the Yellow Report, namely:

> Documentation and preservation of **simulation and reconstruction tools, analysis code, data products and workflows**.
> Documentation and preservation of **data and software** required for detector development, e.g. **test beam experiments**.
> **Catalog of MC data samples** for the design of the reference detector, including the **event generator data as well as full simulations data**.

This will serve well to not only have unprecedented preservation of design motivations and decisions but also to develop policies and expertise well before collisions start, a big advantage over having to scramble after the fact.

The Yellow Report promises

- Documentation and preservation of **simulation and reconstruction tools, analysis code, data products and workflows**.
- Documentation and preservation of **data and software** required for detector development, e.g. **test beam experiments**.
- **Catalog of MC data samples** for the design of the reference detector, including the **event generator data as well as full simulations data**.

This document will start by listing requirements formed from our guiding Statement of Principles as well as from the input of the expert panel. DAP is unusual in the Single Software Stack decision process though, in that meeting requirements goes beyond identifying software solutions and into policies and enforcement much more than previous decisions. We therefore

follow the requirements by Recommendations that do include options for individual issues but also focus on the overarching framework.

Requirements:

The most relevant parts of the EIC Software: Statement of Principles are 4 - User-centered design, 5 - Openness and simplicity, and 6 - Reproducibility. 7, especially regarding wheel reinvention also strongly applies.

Overall:

- "**Holistic approach**": data without the software is useless, as is software without build and verification systems and/or necessary additional data (alignment, calibration, magnetic field maps etc.)
- No matter what preservation tools are developed [...], if they are not conceived **from the beginning** as an integral part of the standard frameworks, retrofitting will be nearly impossible.
- **Access** is as important as archival.
- A data management plan needs to be **exercised** and **tested**.
- Solutions should be **intuitive and easy to use**.
- **Low or no barriers**. The more data, analysis, and document uploads require a gatekeeper's consent, the worse will be the users' uptake
- Prefer **existing solutions**, especially where facility support is needed and exists
- **Discoverability**. All options need to be complemented by catalog tools, search tools, and the like.
- Success relies heavily on users. It requires **training**, as well as **buy-in and enforcement** by conveners and management

Data preservation:

- Test beam data, simulation data, and MCEG data must be **preserved and cataloged**
- Where possible/realistic, they should be **open and discoverable to the public**.

Analysis preservation:

- Plots and tables can be **reproduced by anyone**.

Document preservation:

- Important documents need to be **stored permanently**. Wiki's change.
- We can generate **DOI numbers**.
- Granular (enough) **access controls** to make proprietary information only accessible to the right group of people while making it easy to add and retrieve open documents.

Recommendations and Options:

The EIC CompSW Organization proposes the following DAP policies for the EIC:

1. DAP is recognized as a fundamental component of the EIC research and a factor improving the quality of the research not only in the long, but in the medium and short terms. DAP practices have the potential to improve efficiency and robustness of the design process for the EIC Project Detector and provide an opportunity to design and deploy systems which are DAP-capable over a long period of time. This is especially important since according to the experience in the NHEP community retrofitting DAP on existing systems in the future will be nearly impossible.
2. The collaboration commits to initiating discussions about DAP throughout the leadership structure, from steering committee to working groups.
    a. Identify the software and computing elements of the research conducted in the group which merit preservation. It is expected that at least one area of study in each group will be classified as such.
    b. Package the respective research workflow(s) and tools in a way that ensures reproducibility i.e. captures the software and its configurations, data and other dependencies and the workflow description detailed enough for a qualified person to be able to reproduce the study
3. Development of in-house solutions shall be kept to an absolute minimum in favor of leveraging proven and community-accepted platforms (OpenData, REANA, RIVET, Zenodo, HEPData to name a few).
4. The EIC community shall identify a candidate solution for a digital repository for keeping a variety of materials necessary for preservation. Having such a repository which stores information in a discoverable and persistent manner over a long period of time is an essential component of DAP. Once a platform is chosen, the community will commit to allocating resources for deployment, development and long-term maintenance of the platform.
5. Web documentation is vital for all of the EIC activity. When designing web information services, the long-term durability and ease of maintenance must be addressed. Platforms which require periodic (and potentially disruptive) updates of their components, such as Drupal, should be avoided. Web materials need to be curated by a designated team of editors to combat obsoletion and inefficient layout, and to make sure the materials are optimally accessible.
6. Published papers based on EIC data will provide open access to the data set and analysis workflow with sufficient detail to allow the reader to reproduce the result. In the period preceding the EIC commissioning this may include published Monte Carlo and test beam studies.
7. A EIC-DAP Task Force is formed from EIC members with relevant experience. The Task Force will liaison with the working group conveners and facilities to coordinate the DAP activity. The Task Force will also review the design and deployment strategy of services such as Conditions Database, File Catalog and others, to ensure that they meet the requirements of DAP.

Presenters:

Maxim Potekhin

Ulrich Schwickerath
Ernst Sichtermann

Resources:

https://dphep.web.cern.ch/
https://doi.org/10.5281/zenodo.2653526 “Software Preservation and Legacy issues at LEP” (J.Shiers)
https://arxiv.org/abs/1810.01191 “HSF White Paper: Data and Software Preservation to Enable Reuse”
https://arxiv.org/abs/2203.10057 “2022 Snowmass Summer Study: Data and Analysis Preservation, Recasting, and Reinterpretation”
HERA experiments
https://arxiv.org/abs/2106.11058 "Preservation through modernisation: The software of the H1 experiment at HERA"
https://arxiv.org/abs/1607.01898 “The ZEUS long term data preservation project”
https://inspirehep.net/literature/1211009 “The ZEUS data preservation project”
https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.128.132002 "Measurement of Lepton-Jet Correlation in Deep-Inelastic Scattering with the H1 Detector Using Machine Learning for Unfolding"

Convener Summary:

The discussion made clear that the topic of Data and Analysis preservation is a complex one with many aspects to consider beyond a choice of software packages. These include policy decisions that will require endorsement from the collaboration as a whole and resources to back them up. A task force assigned to this purpose was called for in the discussion and there was consensus that this will be needed. The task force will be organized by two interim co-leads, until the official formation of a collaboration and the role is officially codified. The co-leads will be assigned by the CompSW and SimQA conveners.