"EIC Analysis Knowledge Management"
BNL LDRD-A 2023

Description of the project

This proposal addresses the topic area from the LDRD call: *Research and Development towards the Second Detector and Computing at the Electron-Ion Collider.*

The EIC detector development process involves an extensive simulation effort, which is expected to grow in complexity and sophistication in the next few years. On approximately the same time scale, creation and evolution of analysis tools for EIC is foreseen. Given the projected timeline of the ePIC experiment and the possibility of the second EIC experiment ("Detector 2") on a longer time scale, the issues of effective knowledge preservation, management and transfer come to the forefront. This is confirmed by the experience of the long-running LHC experiments, where onboarding new collaboration members (graduate students and postdocs being one example) proved to be a crucial element of the experiment's lifecycle, while also being a fairly labor-intensive and time consuming process. A closely related but separate issue is **reproducibility** of simulations and analyses, which in the case of the EIC experiments at present time also includes detector optimization and other components of the R&D. In the EIC context, the importance of this work area is further enhanced by the necessity of knowledge transfer from ePIC to the "Detector 2" effort.

At CERN, there are a number of platforms available to the LHC community to achieve these goals, such as the **CERN Open Data** portal, REANA (reproducible analysis platform) and the "CERN Analysis Preservation" (CAP) framework.

Out of these three tools, only one – REANA – has been successfully deployed in the United States, with an instance at BNL SDCC being a prime example. It has been used with success in the PHENIX analysis preservation effort. At present, the EIC community does not have at its disposal a more universal platform such as the CERN Open Data portal, which allows its users to reliably store, query and use in any combination the following components of simulation and/or analysis materials:

- Representative data samples, underpinned by distributed/federated storage
- Software (preserved as code and/or in the form of Docker images or VMs)
- Configuration and conditions data
- Documentation and graphics

Any and all of these elements can be combined to form a cohesive package, which gives the system considerable versatility and makes it a very efficient tool for knowledge preservation and transfer. In addition, its functionality can be leveraged organically to meet the objective of the Open Data policies, and more general goals of data and analysis preservation. In terms of durability of this information, the packages are assigned Digital Object Identifiers (DOIs), which is tremendously helpful in assuring long-term availability of the resources.

The value of this system has been widely recognized internationally and in the United States, as evidenced in active discussions which recently took place in the FAIROS-HEP meeting in 2023. We believe that a system like that has a potential to be a force multiplier in the detector R&D for both ePIC and Detector 2. This proposal builds on the experience gained through active, long-time participation of BNL in the international DPHEP collaboration hosted at CERN and the vigorous analysis preservation effort by the PHENIX Collaboration in 2019-2023. The project would establish active efforts in the following area:

Integrated EIC Analysis Knowledge Management Portal: Taking the existing CERN Open Data portal as an example and a prototype, we shall investigate the possibility of using the open source components of its software with necessary adaptations, to allow for timely development and deployment on the US-based resources. We will complete the design and implementation of this system. The portal will have both Web-based and programmatic interfaces and will contain a well organized and indexed set of information packages encompassing software, data samples, all kinds of documentation, tutorials and any other materials as necessary for a given research topic.

Expected results

A functional Knowledge Management Portal, which will be implemented as a Web service with a user-friendly and efficient UI, as well as a CLI/programmatic interface. It will be targeted initially at EIC Detector 2 but will be developed as an experiment agnostic community-level service.

Funding opportunities & programmatic benefits

The CERN Open Data portal (which serves as an inspiration for this proposal) serves the needs of multiple LHC and other experiments in the HEP/NP area. Likewise, the proposed platform in the US, while being developed primarily for the needs of the EIC community can naturally become a cross-cutting solution for a number of experiments at BNL (such as sPHENIX) and in the more broad context of HEP/NP in the United States. This platform is uniquely suited to fulfill the mandate for Open Data which has become an established policy of the DOE Office of Science. The project has the potential to attract funding for development and support of a community-wide service. It is exceptionally well aligned with the Analysis Preservation agency guidelines and sustainable software initiative in the community.

Personnel

All personnel have backgrounds in distributed computing: Wenaus (ATLAS), Smirnov (STAR), Potekhin (STAR, Belle II, DUNE, PHENIX), Diefenthaler (Nuclear Physics experiments at JLAB). Potekhin has a few years of experience in Data and Analysis preservation in PHENIX and is an active member of the CERN DPHEP collaboration.

PI:
Maxim Potekhin, BNL lead on EIC software infrastructure, PHENIX Analysis Preservation lead

Key personnel:
Dmitri Smirnov, PHENIX Analysis Preservation developer
Torre Wenaus, ePIC Deputy Software and Computing Coordinator

Requested budget

1 postdoc + 30% of existing staff members for 3 years.