Contribution ID: **13** Type: **Poster**

Data Knowledge Base for HENP Scientific Collaborations

Tuesday 22 August 2017 16:25 (20 minutes)

Scientific collaborations operating on modern facilities generate vast volumes of data and auxiliary metadata, and the information is constantly growing. High energy physics data is a long term investment and contains the potential for physics results beyond the lifetime of a collaboration or/and experiment. Many existing HENP experiments are concluding their physics programs, and looking for ways to preserve their data heritage.

The Run1,2 estimated LHC experiments RAW data volume is 15+ PB/year and ˜130 PB/year is expected by the end of Run 3 in 2022 and 200 PB/year for High-Luminosity LHC Run. Even today the managed data volume of the ATLAS experiment is close to 300PB.

Data Preservation HEP working group announced Data Preservation Model in May 2012. This model includes the preservation of real and simulated data, the analysis level, reconstruction and simulation software, and the preservation of documentation (such as internal notes, wikis, etc). However, existing long-term preservation resources are loosely connected and don't provide tools for the automatic reproduction of connections/links between various data and auxiliary metadata from storage subsystems. Data Knowledge Base R&D Project, started in 2016, is aimed at developing of the software environment and providing a coherent view/representation of the basic information preservation objects/components. The present architecture of DKB is based on the ontological model of HENP studies. The central storage –OpenLink Virtuoso –consolidates the basic metadata about scientific papers and internal documents, experimental environment and data samples, used in physical analysis. Specific services –metadata export/extraction/import tools, aggregation/integration modules –are organized as workflows run by Apache Kafka, providing nonstop data processing. One of the most challenging tasks is to establish and keep connectivity between data samples and scientific publications, internal notes and conference talks. Scientific publications and notes have information to establish the above connectivity. (Meta)Information could be extracted from papers and would be used to connect data samples and analysis, and then import these links into Virtuoso in accordance with the ontological model. As a result, all data samples, used in the data analysis described in the document of interest, can be obtained with a simple SPARQL request to Virtuoso. In the nearest future DKB architecture is planned to be enhanced with the SPARQL-endpoint services for InSpire HEP, CERN Document Server and Production Systems, thus providing virtual integration of these storage resources.

Authors: GRIGORYEVA, Maria (Institute for Theoretical and Experimental Physics (RU)); GOLOSOVA, Marina (Institute for Theoretical and Experimental Physics (RU)); WENAUS, Torre (Brookhaven National Laboratory (US)); KLIMENTOV, Alexei (Brookhaven National Laboratory (US))

Presenter: PADOLSKI, Siarhei (BNL)

Session Classification: Poster Session

Track Classification: Track 1: Computing Technology for Physics Research