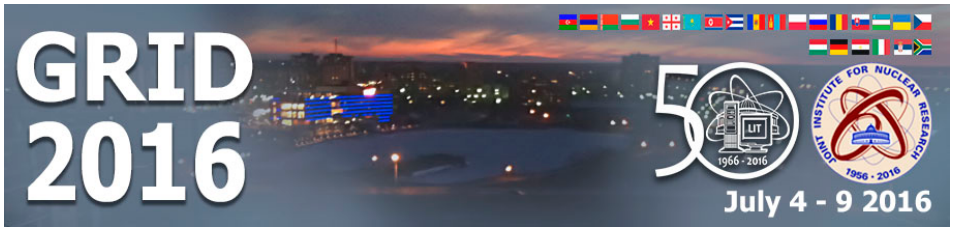

Contribution ID: **89** Type: **Sectional reports**

Data Knowledge Base Prototype for Collaborative Scientific Research

Thursday 7 July 2016 09:20 (20 minutes)

The most common characteristics of large-scale modern scientific experiments are long lifetime, complex experimental infrastructure, sophisticated data analysis and processing tools, peta- and exascale data volume. All stages of an experiment life cycle are accompanied with the auxiliary metadata, required for monitoring, control and scientific results replicability and reproducibility. The actual issue for the majority of scientific communities is a very loose coupling between metadata describing data processing cycle, and metadata representing annotations, indexing and publication of the experimental results. Researchers from Kurchatov Institute and Tomsk Polytechnic University have investigated main metadata sources for one of the most data-intensive modern experiment - ATLAS, at LHC. It has been noticed that there is a lack of connectivity between data and meta-information, for instance between physics notes and publications, and data collection(s) used to conduct the analysis. Besides, to reproduce and to verify some previous data analysis, it's very important for the scientists to mimic the same conditions or to process data collection with new software releases or/and algorithms. That's why all information about data analysis process must be preserved, starting from the initial hypothesis following by processing chain description, data collection, initial results presentation and final publication. A knowledge-based infrastructure (Data Knowledge Base - DKB) gives such possibility and provides fast access to relevant scientific and accompanying information. The infrastructure architecture has been developed and prototyped. DKB is functioning on the basis of formalized representation of scientific research lifecycle –HEP data analysis ontology. The architecture has two data storage layers: Hadoop storage, where data from many metadata sources are integrated and processed to obtain knowledge-based characteristics of all stages of the experiment, and Virtuoso ontology database, where all extracted data are registered. DKB agents process and aggregate metadata from data management and data processing systems, metadata interface, conference notes archives, workshops and meetings agendas, and publications. Additionally, this data is linking with the scientific topic documentation pages (such as Twikis, Google documents, etc) and information, extracted from full texts of experiment supporting documentation. In this way, rather than require the physicists to annotate all meta-information in details, DKB agents will extract, aggregate and integrate all necessary metadata automatically. In this talk we will outline our accomplishments and discuss the next steps and possible DKB implementation in more details.

Author: Ms MARIA, Grigorieva (NRC KI)

Co-authors: Mr ALEKSEEV, Aleksandr (Tomsk Polytechnic University); Dr KLIMENTOV, Alexei (Brookhaven National Lab); Mr GUBIN, Maksim (Tomsk Polytechnic University); Ms GOLOSOVA, Marina (National Research Center "Kurchatov Institute"); Ms OSIPOVA, Victoria (Tomsk Polytechnic University)

Presenter: Ms MARIA, Grigorieva (NRC KI)

Session Classification: Plenary reports

Track Classification: 9. Consolidation and integration of distributed resources