20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)

Contribution ID: **216** Type: **Oral presentation to parallel session**

Data and Software Preservation for Open Science (DASPOS)

Monday 14 October 2013 17:48 (20 minutes)

Data and Software Preservation for Open Science (DASPOS), represents a first attempt to establish a formal collaboration tying together physicists from the CMS and ATLAS experiments at the LHC and the Tevatron experiments with experts in digital curation, heterogeneous high-throughput storage systems, large-scale computing systems, and grid access and infrastructure. Recently funded by the National Science Foundation, the project is organizing multiple workshops aimed at understanding use cases for data, software, and knowledge preservation in High Energy Physics and other scientific disciplines, including BioInformatics and Astrophysics. The goal of this project is the development and specification of an architecture for curating HEP data and software to the point where the repetition of a physics analysis using only the archived data, software, and analysis description is possible. The novelty of this effort is this holistic approach, where not only data but also software and frameworks necessary to use the data are part of the preservation effort, making it true “physics preservation”rather than merely data preservation. This effort is an exploration of the problems to be solved at the technical, sociological, and policy levels, in order for integrated data preservation as envisioned by DPHEP to be possible. This work will provide the solid foundation necessary for the next steps of preservation infrastructure development. The research is a combination of two overlapping activities: a “horizontal” coordination and consensus-forming activity, both internal to HEP and including other disciplines, to agree on prototype metadata definitions and other common aspects of data preservation, and the more technical construction of the “vertical”slice of archival infrastructure. One measure of success, the so-called “Curation Challenge”will be a small-scale but full system test of a particular archiving solution enabling the discovery and enumeration of the critical issues in establishing preservation architectures. A key aspect of this work will be the inclusion of different scientific disciplines in the discussions of research use cases, archival strategies, metadata definitions, and policy considerations. Through this extended dialogue, we will be able to establish elements of commonality that can lead to shared technical and architectural solutions across disciplines. We also expect to outline branch points throughout the preservation architecture specification where policy choices will dictate technical outcomes, leading to a blueprint for any discipline approaching the problems of large-scale data preservation and open access. We aim for these common solutions and principles established to serve a similar role within the HEP community that the OAIS (Open Archival Information System) model plays for Trusted Digital Repositories. Of equal importance to these broad-ranging policy and technology issues will be the training of a team of graduate students in the technical aspects of large data set preservation, global grid-based access tools, and other facets of this multi-disciplinary problem. Finally, the development of technologies for the preservation of large scientific data archives opens up the possibility of future scientific opportunities and insights not otherwise available.

Primary authors: HILDRETH, Mike (University of Notre Dame (US)); HILDRETH, Mike (Department of Physics-College of Science-University of Notre Da)

Co-authors: WATTS, Gordon (University of Washington (US)); BLOOM, Kenneth (University of Nebraska (US)); NEUBAUER, Mark (Univ. Illinois at Urbana-Champaign (US)); NEUBAUER, Mark Stephen (Univ. Illinois at

Urbana-Champaign); GARDNER JR, Robert William (University of Chicago (US))

Presenter: HILDRETH, Mike (University of Notre Dame (US))

Session Classification: Data Stores, Data Bases, and Storage Systems

Track Classification: Data Stores, Data Bases, and Storage Systems