Contribution ID: **405** Type: **oral presentation**

Open Data and Data Analysis Preservation Services for LHC Experiments

Tuesday, April 14, 2015 2:00 PM (15 minutes)

In this paper we present newly launched services for open data and for long-term preservation and reuse of high-energy-physics data analyses. We follow the “data continuum” practices through several progressive data analysis phases up to the final publication. The aim is to capture all digital assets and associated knowledge inherent in the data analysis process for subsequent generations, and to make a subset available rapidly to the public.

A data analysis preservation pilot study was launched in order to assess the usual workflow practices in LHC collaborations. Leveraging on synergies between ALICE, ATLAS, CMS and LHCb experiments, the analysed data was followed through various “analysis train” steps, from the initial capture and pre-selection of primary data, through several intermediate selection steps yielding more greatly reduced datasets, up to the final selection of N-tuples used for producing high-level plots appearing in scientific journal publications. Most of the analysis chain is kept strictly restricted within a given collaboration; only the final plots, presentations and papers are usually made public. It is therefore essential to handle access rights and embargo periods as part of the data life cycle.

The study revealed many similarities between collaborations, even though the variety of different practices existing in different groups within the collaborations make it hard to reproduce an analysis at a later time in a uniform way. One recurring problem underlined by the study was to ensure an efficient “knowledge capture” related to user code when the principal author of an analysis (e.g. a PhD student) leaves the collaboration later.

The pilot solution has been prototyped using the Invenio digital library platform which was extended with several data-handling capabilities. The aim was to preserve information about datasets, the underlying OS platform and the user software used to study it. The configuration parameters, the high-level physics information such as physics object selection, and any necessary documentation and discussions are optionally being recorded alongside the process as well. The metadata representation of captured assets uses the MARC bibliographic standard which had to be customised and extended in relation to specific analysis-related fields. The captured digital assets are being minted with Digital Object Identifiers, ensuring later referencability and citability of preserved data and software. Connectors were built in the platform to large-scale data storage systems (CASTOR, EOS, Ceph). In addition, to facilitate information exchange among concerned services, further connectors were built to

the internal information management systems of LHC experiments (e.g. CMS CADI), to the discussion platforms (e.g. TWiki, SharePoint), and to the final publication servers (e.g. CDS, INSPIRE) used in the process. Finally, the platform draws inspiration from the Open Archival Information System (OAIS) recommended practices in order to ensure long-term preservation of captured assets.

The ultimate goal of the analysis preservation platform is to capture enough information about the process in order to facilitate reproduction of an analysis even many years after its initial publication, permitting to extend the impact of preserved analyses through future revalidation and recasting services.

A related “open data” service was launched for the benefit of the general public. The LHC experimental collaborations are committed to make their data open after a certain embargo period. Moreover, the collaborations also release simplified datasets for the general public within the framework of the international particle physics masterclass program. The primary and reduced datasets that the collaborations release for public use are being collected within the CERN Open Data portal service, allowing any physicist or general data scientist to access, explore, and further study the data on their own.

The CERN Open Data portal offers several high-level tools which help to visualise and work with the data, such as an interactive event display permitting to visualise CMS detector events on portal web pages, or a basic histogram plotting interface permitting to create live plots out of CMS reduced datasets. The platform guides high-school teachers and students to online masterclasses to further explore the data and improve their knowledge of particle physics. A considerable part of the CERN Open Data portal was therefore devoted to attractive presentation and ease-of-use of captured data and associated information.

The CERN Open Data platform not only offers datasets and live tools to explore them, but it also preserves the software tools used to analyse the data. It notably offers the download of Virtual Machine images permitting users to start their own working environment in order to further explore the data; for this the platform uses CernVM based images prepared by the collaborations. Moreover, the CERN Open Data platform preserves examples of user analysis code, illustrating how the general public could write their own code to perform further analyses.

Primary authors: COWTON, Jake (University of Northumbria at Newcastle (GB)); KUNCAR, Jiri (CERN); RUEDA GARCIA, Laura (CERN); FOKIANOS, Pamfilos (National and Kapodistrian University of Athens (GR)); HERTERICH, Patricia Sigrid (Humboldt-Universitaet zu Berlin (DE)); DALLMEIER-TIESSEN, Sunje (Humboldt-Universitaet zu Berlin (DE)); Dr SIMKO, Tibor (CERN); SMITH, Tim (CERN)

Presenter: SMITH, Tim (CERN)

Session Classification: Track 5 Session

Track Classification: Track5: Computing activities and Computing models