

ALICE Data Preservation plans

The ALICE collaboration is committed to develop a long term program for Data Preservation to serve the triple purpose of i) preserving data, software and know-how inside the Collaboration, ii) sharing data and associated software and documentation with the larger scientific community, and iii) give access to reduced data sets and associated software and documentation to the general public for educational and outreach activities.

The long-term strategy to achieve these goals together with the scope of Data Preservation plan have been approved by the Collaboration. The implementation of the global strategy is still under development and subject of discussion between the ALICE Offline project and the ALICE Physics Board.

As a first step, ALICE has developed an elementary demonstrator aimed exclusively at educational purposes within the CERN Master Class program. The demonstrator features several physics examples of data analysis exercises including a simplified event display. The analysis (V_0 decays of strange particles, inclusive transverse momentum spectrum, nuclear modification factor) are applied to a limited set of data adequately selected among the data collected during the LHC RUN1 period for pp, p-Pb and Pb-Pb collisions.

The analyses are integrated in a CernVM virtual machine hides the complexity of the ALICE software and its installation requirements. The demonstrator is deployed in the open data DPHEP portal at CERN, developed by the IT department. This out of the box ready to run solution provides access to the data, to the software tools and to the documentation needed to perform the analysis and does not require any environment setting or compilation.

The next step is extending the virtual machine concept to provide to ALICE collaborators the long-term access to data and the associated software versions. The latter includes the operating system, software framework, condition data, analysis macros and documentation. The goal is to require the reproducibility of analysis in such virtualized environment as a prerequisite for publishing results. By automating the process of versioning and recording the versions of OS, conditions data and software we will allow any collaborator to reproduce at any time on any hardware platform any published results and to repeat the analysis with new inputs within the environment in which the analysis has originally been performed.

This concept will be first exercised within the ALICE collaboration and then gradually make more and more data and analysis examples will be made available to the general public using exactly the same technology and building on Open Data sharing platforms such as DPHEP portal at CERN.