ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of investment in human and financial resources by the international community. These data embed unique scientific information for the in depth understanding of the profound nature and origin of matter. Because of their uniqueness, long term preservation must be an essential objective of the data processing framework and will lay the foundations of the ALICE Collaboration legacy to the scientific community as well as to the general public. These considerations call for a detailed assessment of the ALICE data preservation strategy and policy. Documentation, long term preservation at various levels of abstraction, data access and analysis policy and software availability constitute the key elements of such a data preservation strategy allowing future collaborators, the wider scientific community and the general public to analyze data for educational purpose and for eventual reassessment of the published results. The present document describes the basic principles that will guide the redaction addressed by the ALICE data preservation policy.

ALICE data formats

The level of abstraction of ALICE data increases at every step of the data processing chain starting from basic raw data delivered by the detectors of the experiment evolving into physics-analysis-ready data and ending with physics data suitable for publication. At each stage of the data processing ancillary meta-data, such as calibration, alignment and running condition parameters are invoked to transform raw detector information into physics information, free of detector biases. The various ALICE data formats are classified as follows:

a) *Raw data* embedding the signals delivered by the detectors along with the associated status data containing various information on the running conditions constitutes the primary information collected by the ALICE experiment. They provide the input of the reconstruction algorithm, together with the calibration data stored in a dedicated database;

b) *Monte-Carlo data*, including data at the event generator level (MC truth) and data mimicking the raw data format (digits), anchored to real data reproducing the running conditions;

c) *Event Summary Data* (ESD) produced by the reconstruction algorithms, for both Monte Carlo and raw data. The ESD events provide calibrated tracks in a generic format, but also additional detector specific information allowing a full physics analysis;

d) General purpose *Analysis Object Data* (AOD), derived from ESD data. The AOD data format contains a simplified event model with few additional high level detector specific parameters;

e) *Custom analysis object data*, used standalone or together with the general purpose AOD for specific analysis;

f) *Published physics results* and highly abstracted data resulting from the analysis.

These different formats of the ALICE data lead to a specific schema for data preservation. While formats can change with time, the collaboration provides software releases suitable to read and process any format, or alternatively to migrate data from one format to another. Since processed data can exist in several versions, only the version used for the final publication of the results is considered as a candidate for data preservation.

The ALICE Computing Model includes the provision for permanent storage of two copies of the raw data. They are not presently being considered for open access, but they can be reprocessed at any time by members of the ALICE collaboration upon approval by the ALICE Physics Board. The original datasets used to produce published results, together with the adequate software version (framework and macros) are subject to long-term preservation.

The data of ALICE at all levels is stored in non-proprietary formats (ROOT[1]); these formats are subject to documented schema evolution due to intrinsic analysis needs. This is a highly controlled process under the supervision of the ALICE Physics Board, following ROOT rules. The changes in the software following such an evolution are ensured to be backward compatible. Major changes in the data formats, due for example to the upgrade of the detector or to other legitimate reasons can generate non backward compatible changes, requiring the preservation of the initial software version. In addition, such changes require us either to provide modules allowing one to read the data transiently into the new format, or to run filtering procedures to migrate the old data sets into the new format.

ALICE provides long term support for preserving the software used to process all the above-mentioned data, including the exhaustive documentation of the procedures. The main goals of the long term preservation are:

- The reproducibility of published results by the collaboration;
- Providing for the re-processing at any stage of the data chain, with the possibility to modify the procedure;
- Transparency in the analysis procedures allowing the re-analysis of data by third party categories.

Open access

The ALICE collaboration agrees with the principle of open access to data, software and documentation, that will allow the processing of data by non-ALICE members under the conditions listed in the future policy implementation document. Data with high abstraction, such as AOD, will be conditionally made publicly available after an embargo period of 5 years after publication for 10% of the data and 10 years for 100% of the data. Depending on the available resources for open access, external users can be conditionally granted access to computing resources to process data. Different levels for preservation and open access are currently foreseen:

Level 1: Published results and related data (analysis notes, numerical data, figures)

All ALICE scientific results are public. They are published in international scientific journals adopting an open access policy. The numerical results, procedures and intermediate meta-data are made available through publications on trusted third-party platforms such as *Inspire* or *HEPData*.

Level 2: Simplified data formats for analysis

The simplified data formats are derived from the AOD and contain selected highly abstracted data such as: energy-momentum four-vectors, particle identification or centrality. For outreach and educational purposes, limited data sets will, under conditions, be made publicly available along with the associated software. Due to the high level of abstraction, these data can be processed by non-ALICE software and may be stored in third-party repositories according to a release policy.

Level 3: Reconstructed data and Monte Carlo data, together with analysis software

ALICE stores the reconstructed data on transient storage. These data are available for the ALICE members in two forms, serving different use cases: while the ESD format is mostly used for calibration and detector studies, the AOD format is internally used for analysis. Both of these formats are produced after reconstructing real and simulated data. The ALICE data preservation model foresees making exclusively the AOD format and the Monte Carlo truth data publicly available. In addition only a fraction of the data will be made available on a time scale of 5 years for 10% of the data and 10 years for 100% of the data.

Together with these subsets, ALICE will also make available the software needed to process these data and the existing documentation. Depending on the available resources, non-ALICE members will be given the possibility to access these data under well-controlled conditions using a dedicated computing infrastructure. Any publication that results

[1] http::root.cern.ch

from data analysis by non-members of the collaboration will require a suitable acknowledgement ("data was collected by ALICE") and disclaimer ("no responsibility is taken by the ALICE collaboration for the results published here").

Level 4: Raw data and the associated software

The raw data processing stage requires a detailed knowledge of the ALICE detector and extensive computing resources. ALICE stores the original raw data on Tier-0 tape storage, plus a custodial copy on Tier-1 disk storage as secured backup. All versions of ALICE reconstruction software are required to be able to process raw data sets from the beginning of the data taking, allowing the data reprocessing in successive passes, as the calibration gets refined. This procedure allows the long-term preservation of the data and software. ALICE does not currently consider these data suitable for the general public but leaves open the possibility of re-processing, by members of the collaboration and after approval by the ALICE Physics Board.

Data release policy and responsibilities

For the widest possible re-use of the data, while protecting the Collaboration's liability and reputation, data could be released under the emerging standard Creative Commons CC0 waiver[2]. The license policy will be detailed in the implementation document. Data will also be identified with persistent data identifiers, and it is expected that the third parties cite the public ALICE data through these identifiers, so that its re-use can be monitored and contribute to the assessment of the impact of the LHC program.

This data preservation, re-use and access activity implies responsibilities across several bodies within and beyond the collaboration. The Collaboration Board will approve, uphold and possibly amend this policy. The responsibility for the implementation phase of this policy related to setting up the infrastructure and procedures, can be delegated to a Data Preservation Coordinator. The ALICE offline collaboration, under the supervision of the DP Coordinator, will be responsible for the management of resources allocated for public data releases and of the infrastructure allowing re-use of these data. The Data Preservation Coordinator will propose to the Collaboration Board the release of data for wider access. The proposal will specify quality, quantity and location of the data to be released. The Coordinator will assure that the policy is followed, within the plan and resources approved by the management board for this purpose.

[2] http://creativecommons.org/publicdomain/zero/1.0 . The CC0 license is designed for data, and allows re-use by anyone, under the responsibility of these final users. The emerging standard practice in disciplines where data reuse is common expects that third parties cite the original author of the data (e.g. through DOI, Digital Object Identifiers, soon available through INSPIRE).