Open Data implementation in ALICE: Status & Plan

Ana Marin,
Alexandru Florin Dobrin, Stefano Piano
on behalf of ALICE

CERN Open Data Policy

- Endorsed by ALICE Collaboration Board in November 2020
- The policy commits to publicly releasing level 3 scientific data:
  - Input to most physics studies (AOD or derived data formats)
  - To be released alongside the software and documentation needed to use the data
  - Allowing high-quality analysis (needed to be also released MC AOD)
- Public data releases expected periodically
- Needed appropriate latency period to allow:
  - thorough understanding of the data
  - reconstruction and calibrations
  - the scientific exploitation of the data by the collaboration
- Aim to commence data releases **within five years** of the conclusion of the run period
- Size of the released datasets commensurate with the amount of data collected
- Full datasets will be made available at the end of the collaboration

ALICE plans for Open Data Implementation

- Set up CERN Open Data Portal with sample of ALICE data(OD) Portal:
  - Current status: 5% (7%) of Pb-Pb (pp) 2010 ESD datasets released, totaling 6.5 TiB
  - Preparation of Run 1 and Run 2 data and simulated data with the new data format:
    - Run 3 AOD (AO2D and MCAO2D)
    - Open Data Quality Control
- (Simple/Run 3) ALICE analysis demonstrator in CERN Open Data portal
  - As already done for Run 1 & Run 2 AliPhysics Analysis framework:
    - VM and docker container to ease portability
    - Integration with REANA to run analysis directly on Open Data
- Documentation

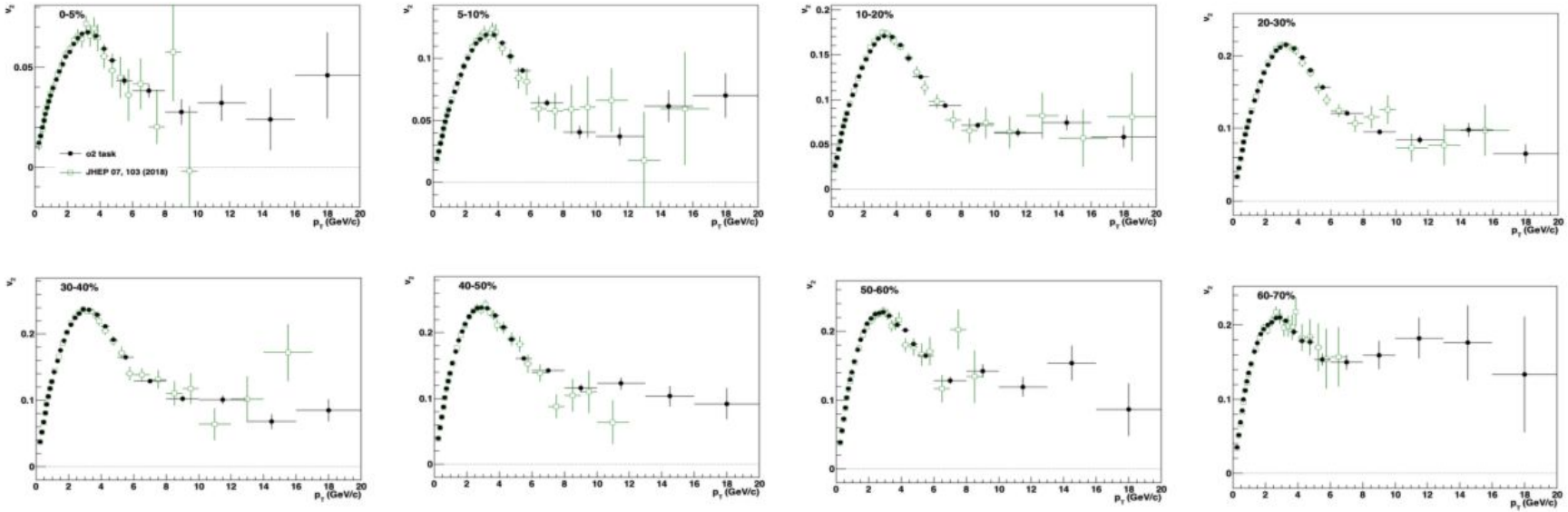
New ALICE Open Data Format

- New AO2D format to be published as open data
- Based on the new data format and software framework developed by ALICE O2 project for Run 3 and Run 4:
  - It will ensure data preservation of Run 1 and Run 2 data
  - Significantly reduced size per collision (factor 16 wrt Run 2 ESD and factor 5 wrt Run 2 AOD)
  - New flat data model optimized for fast IO (>10x faster than Run 2 AOD)
  - Possible to adopt (skimmed) derived data set like nanoAOD format to compress further
- Such a refurbishment required a long conversion production of all Run 1 and Run 2 ESDs and AODs into new AOD format for both data and MC
  - Conversion was done in 2022, but the new data format required additional efforts to make the new analysis framework work with the old data in standalone mode (with a reasonable size)
  - All the Run 1 and Run 2 results have been published with the old analysis framework

New ALICE Open Data Analysis Framework

- Standalone compilation of the Run 3 OD ALICE Analysis Framework:
  - Required some efforts of the offline group to enable standalone local compilation
  - Now available at https://github.com/AliceO2Group/O2OpenAccess
  - Documentation for compilation available on README.md
- Fully working, possible to run the analysis on MCAO2D and AO2D:
  - Provided some examples of analysis tasks
    - Flow task for unidentified particles and PID, MC reconstruction efficiency
- Still some open points before publishing data:
  - PID calibration + centrality estimators
    - Not possible to import them from the old framework
    - They are being regenerated for all Run 1 and Run 2 runs
    - Incorporation in database snapshot underway
  - Data validation

Validating the flow task: v_2 vs p_T



- Good agreement with published data at 5.02 TeV

Expected Open Data release in the next years

- Following the implementation document ALICE plans to release:
  - pp, p-Pb and Pb-Pb Run 1 data samples by the end of 2023 with the new data format
  - 10% of Run 2 data by 2024
    - ALICE will gradually reach 50% of Run 2 data by 2028
  - In 2030 ALICE will start releasing 10% of Run 3 data
- Based on the new AO2D format data volume, we expect:
  - 2023: 35 TB of which 15 TB were planned to be published in 2022 (Run 1)
  - 2024: 105 TB (10% Run 2)
  - 2025 - 2028: 105 TB/year to reach 50% of Run 2 in 2028
- In Run 2 ALICE inspected ~1 nb^{-1} Pb-Pb data, while for Run 3 & Run 4 ALICE plans to collect 13 nb^{-1} of Pb-Pb collisions
  - In 2030 we will start releasing Run 3 data
    - ~2 PB/year publishing AO2Ds ⇒ crucial to publish skimmed derived data sets instead

Summary & Outlook

- ALICE Open Data benefit from the new Analysis Framework improvements:
  - New Run 3 AOD format suitable for Run 1 and 2 Open Data
  - Skimmed derived data sets will provide
    - Further data compression crucial to make public Run 3 and 4 data
    - Higher event throughput and reduction of needed CPU wall time
- Dedicated human resources for the ALICE Open Data have been allocated (2FTEs):
  - ALICE committed to publicly releasing level 3 scientific data
  - Organization of new ALICE Open Data working group
  - ALICE aims at making public Run 1 and Run 2 Open Data from 2023
- Still a few more steps before making public ALICE data with the new format