
# Analysis and data preservation in ALICE

Stefano Piano
on behalf of ALICE


# Introduction

# Analysis and data preservation in ALICE

- Set up ALICE Open Data Portal with sample of our data and analysis
- Simple ALICE analysis demonstrator in CERN Open Data portal
- Now containerized with Docker to ease portability

Current status:

- Obviously most of ALICE efforts focused on the upgrade for RUN3
- Additional dedicated resources needed


# Open Data

# ALICE open data policy (Run1 and Run2)

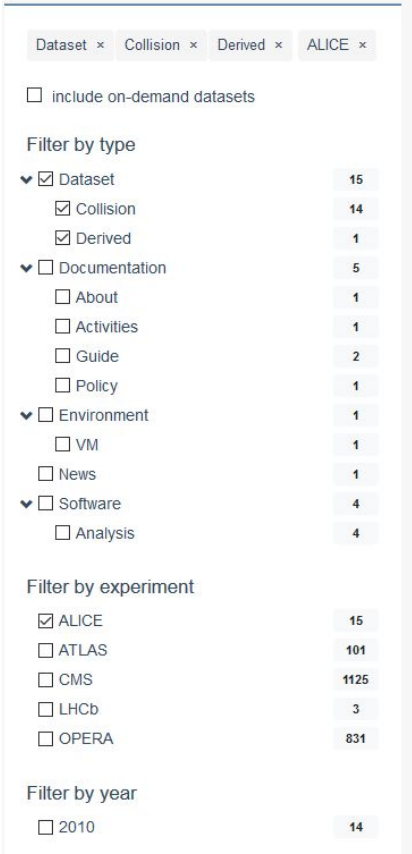
- Different levels for preservation and open access are currently foreseen:

1. All ALICE scientific results are public
2. Simplified data formats for analysis (AOD): for outreach and educational purposes, limited data sets will, under conditions, be made publicly available along with the associated software
3. Reconstructed data and Monte Carlo data, together with analysis software: exclusively the AOD format and the corresponding Monte Carlo data publicly available on a time scale of 5 years for 10% of the data and 10 years for 100% of the data
4. Raw data and the associated software not suitable for the general public

- Document available at http://opendata.cern.ch/record/412

# Open Data (.cern.ch)

- CERN IT platform to share data and software
- Currently published ALICE data
  - only sample from 2010 dataset
  - 14 reconstructed **ESD datasets** (Minimum Bias interactions)
  - LHC10b 7 TeV pp collisions (a few files for Masterclasses)
  - LHC10c 7 TeV pp collisions $27 \cdot 10^6$ (**7%** of $400 \cdot 10^6$)
  - LHC10h 2.76 TeV PbPb collisions $2.9 \cdot 10^6$ (**5%** of $53 \cdot 10^6$)
- AOD not published yet (only for Pb-Pb 2010 ~40 TB)




# Data Preservation

# ALICE Data Preservation Strategy

- Purpose of ALICE data preservation
  - Preserve data and software inside ALICE for long term use (... beyond the ALICE lifetime)
  - Give access to reduced datasets to the general public for educational and outreach activities

- For long term preservation use:
  - **Open Data, web portal provided by CERN IT**
  - **CAP - CERN Analysis Preservation**

# Analysis Preservation

- Re-use inside ALICE: rerun analysis trains
- Few tests to preserve analysis steps after the trains
- LEGO trains create JSON file for each train run:
  - Describe the data analysis process
  - User can trigger the transfer of JSON to **CERN Analysis Preservation**
  - Additional information can be added afterwards on CAP manually
  - If LEGO trains are not used, the full entry has to be generated manually

- Work on a CAP entry with multiple people
- Add local macros in the CAP web page
- Share finished entry with the whole collaboration

# Information to Preserve (schema still incomplete)

- Used dataset
  - Identifier in Run Condition Table
  - Run numbers
- Computing infrastructure
  - ALICE analysis configuration
- Analysis code
  - AliPhysics code on Github
  - AddTask in AliPhysics
  - Code configuration
  - LEGO train run
  - plotting macros
- Link to documentation/publications
  - ALICE analysis note
  - Journal reference

Information from the LEGO trains

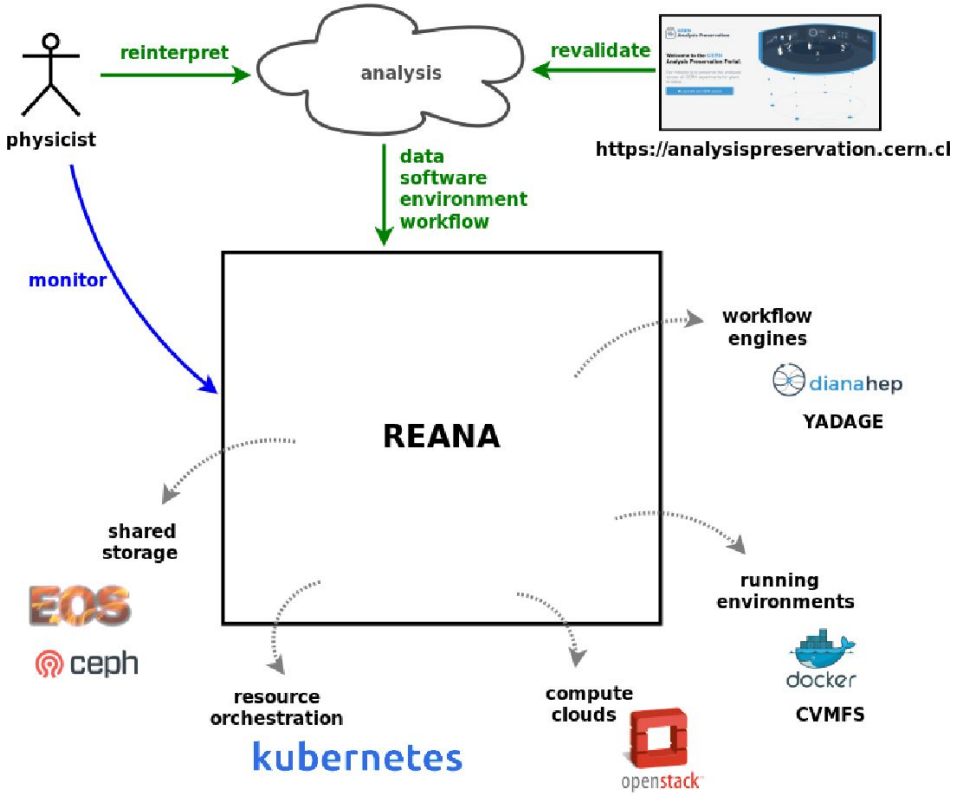
RCT and Github repository have to be preserved separately

To be added manually, but it could be extracted from Monalisa


# Reproducible Analysis

# REusable ANAlysis

- Plan to use REANA
  - Data: open data
  - Use input from CAP
  - Software: CVMFS
  - Environment: LEGO trains
- Can be used for:
  - Rerun the train
  - Plot production with local macros
- Procedure is available on the REANA main page:
  - to run a lego train in a AliPhysics specific container
  - use as input the intermediate files of the lego system, not yet any pre-saved JSON configuration file
- In addition, simple ALICE analysis demonstrator submitted to the CERN Open Data portal works with ALICE Open Data VM, now available in REANA with docker container



ALICE

# Summary & Outlook

- Upgrades for Run3 preparation are currently binding resources
- Data Preservation with Open Data:
  - ~5% of 2010 data sample
  - ~40 TB of free disk space to publish partially 2010 and 2011 AOD's
  - Find more storage capacity for the other datasets
- Demonstrated feasibility to integrate the Analysis Preservation into the ALICE Run2 analysis workflow:
  - Preserve procedure to (re)create approved plots or analyses

- Present procedure should be revised in the context of new Run3 software framework