Preserving and reusing high-energy-physics data analyses

S. Dallmeier-Tiessen\textsuperscript{2}, R. Dasler\textsuperscript{2}, P. Fokianos\textsuperscript{2}, J. Kunčar\textsuperscript{1}, A. Lavasa\textsuperscript{2}, A. Mattmann\textsuperscript{2}, D. Rodríguez\textsuperscript{1}, T. Šimko\textsuperscript{1}, A. Trzcinska\textsuperscript{2}, I. Tsanaktsidis\textsuperscript{2}

\textsuperscript{1}CERN Information Technology
\textsuperscript{2}CERN Scientific Information Service

Open Repositories 2017 · Brisbane, Australia · 26–30 June 2017
Long-term value of data!

Collaborations publish papers even $\sim 15$ years after data taking ends.

JADE data (1979–1986) still unique even $\sim 35$ years later.

Achim Geiser  https://indico.cern.ch/event/588219

DPHEP  https://arxiv.org/abs/1205.4667
Long-term value of knowledge?

CMS collaboration

Experimental physics done by groups of $\sim 3000$ physicists.

Career after PhD

THE ROYAL SOCIETY

High turnover of young researchers.
CERN Analysis Preservation

- A platform for **preserving knowledge** and **assets** of an individual physics analysis.
- Capturing the elements needed to **understand** and **rerun** an analysis even several years later:
  - ✓ data
  - ✓ software
  - ✓ environment
  - ✓ workflow
  - ✓ context
  - ✓ documentation

- Advanced **search** for high-level physics information
- Applying standard **collaboration access restrictions**

*Developed by CERN IT and CERN SIS in close collaboration with LHC experiments*
System overview

![Diagram illustrating the system overview process](https://analysispreservation.cern.ch/)

**ANALYSIS**
- physicist
- analyse
- software
- workflow
- data
- environment

**1 ORGANISE**
- search
- retrieve
- reinterpret

**2 CAPTURE**
- push
- pull

**3 REUSE**
- push
- pull

**COLLABORATION INTERNAL TOOLS AND DATABASES**
- ALICE
- ATLAS
- CMS
- LHCb

@tiborsimko
1. Describing an analysis

Structuring knowledge behind research data analysis.

- JSON Schema
- W3C DCAT
- domain-specific fields
2. Capturing an analysis

Taking consistent snapshot of analysis assets at a certain time.

- datasets: local storage, cloud storage
- software: Git, SVN
- information: DBs, TWiki, SharePoint
- protocols: HTTP, XRootD

@tiborsimko
3. Reusing an analysis

Instantiating preserved analysis on the cloud.

https://analysispreservation.cern.ch
REANA = RE usable ANALyses

- a system for **reusable analysis** execution on the cloud
  - [https://reanahub.io](https://reanahub.io)

- supporting **multiple scenarios**
  - multiple computing clouds
    - CERN OpenStack
  - multiple running environments
    - Docker with CVMFS
  - multiple resource orchestration
    - Kubernetes
  - multiple workflow engines
    - Yadage
  - multiple shared storage systems
    - Ceph, EOS

- close **collaboration** with DASPOS and recast
REANA is FOSS

REANA - Reusable Analyses

Navigation

1. Introduction
   - 1.1. About
   - 1.2. Features
2. Installation
   - 2.1. Installing REANA client
   - 2.2. Installing REANA cloud
   - 2.3. Configuring cluster
   - 2.4. Initialising cloud
3. Getting started
   - 3.1. About
   - 3.2. Install minikube
   - 3.3. Start minikube
   - 3.4. Install REANA
   - 3.5. Initialise REANA cloud
   - 3.6. Run "hello world" example application
   - 3.7. Run "word population" example analysis
   - 3.8. Washing our bowl
4. Examples
   - 4.1. Hello world
   - 4.2. Jupyter notebook
   - 4.3. ROOT and RooFit
5. Architecture
   - 5.1. Overview
   - 5.2. Technology

REANA is a system that permits to instantiate research data analyses on the cloud. It uses container-based technologies and was born to target the use case of particle physics analyses in LHC collaborations. The system paves the way to reusing and interpreting preserved data analyses even several years after the original analysis.
Four questions

1. Input data
   - What is your input data?
     - input files
     - live DB calls

2. Analysis code
   - Which code analyses it?
     - Jupyter notebook
     - custom code

3. Compute environment
   - What is your environment?
     - operating system
     - software & libraries

4. Analysis workflow
   - Which steps did you take?
     - single command
     - complex workflows
Simple example: Jupyter

1. **input**: CSV file

   ```
   FROM centos:7
   RUN yum install -y openssl-release
   RUN yum install -y 
     g++ 
     python-devel 
     python-pip
   RUN pip install ipython==5.0.0 jupyter=1.0.0
   ADD world_population_analysis.ipynb /code/
   ADD World_historical_and_predicted_populations_in_percentage.csv /code/
   WORKDIR /code
   CMD ["jupyter", "nbconvert","world_population_analysis.ipynb"]
   ``

2. **code**: Jupyter notebook

3. **environment**: CentOS7, IP5

4. **workflow**: jupyter nbconvert

   ![Histogram](https://github.com/reanahub/reana-demo-worldpopulation)

   ```python
   In [6]:
   def histogram_by_region(region):
       local_pop = pop[["Region", str(region)]].groupby("Region").sum()
       plot = local_pop.plot(kind="bar", legend=None, title="Percentage of World Population over time in ", str(region))
       plot.set_ylabel("% of world population")
       plot.set_xlabel("")
   In [7]:
   histogram_by_region("Africa")
   ``
Complex example: DAG workflows

- **case studies** in high-energy-physics with LHC collaborations
  - ALICE AliPhysics post-LEGO train analysis
  - ATLAS multi-B-jets analysis
  - LHCb Lb2LcD0K analysis and data production

- **yadage** parametrised workflow engine

Lukas Heinrich  http://github.com/diana-hep/yadage
Reusability ⇄ Preservation

- Reuse
- Preserve
- Use
- Archive

ReANA

CERN Analysis Preservation

@tiborsimko
Conclusions

CERN Analysis Preservation

🔗 http://analysispreservation.cern.ch
🔗 http://github.com/cernanalysispreservation
✉️ analysis-preservation-support@cern.ch

Invenio

🔗 http://inveniosoftware.org
🔗 http://github.com/inveniosoftware
🐦 inveniosoftware
✉️ info@inveniosoftware.org

REANA

🔗 http://reanahub.io
🔗 http://github.com/reanahub
🐦 reanahub
✉️ info@reanahub.io