Preserve to reuse

Tibor Šimko

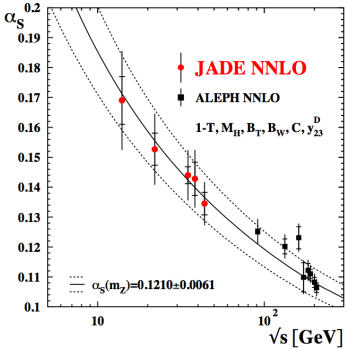
CERN

Snowmass Computational Frontier Workshop

10–11 August 2020

https://indico.fnal.gov/event/43829

Preserve to reuse: a real-life use case

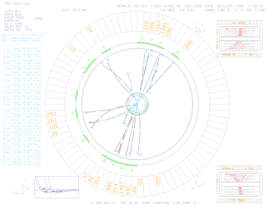


DPHEP `https://arxiv.org/abs/1205.4667`

JADE e^+e^- experiment data (1979–1986) revived after more than twenty years

JADE software porting quest in detail

- Original sources in FORTRAN IV (1974), FORTRAN 77, Sheltran, Mortran, Assembler ported in 2003 to AIX4.3@RS6000, the old routines replaced with CERNLIB libraries and compiled with IBM FORTRAN.
  Huge work!



- In 2017 everything is problematic: AIX@ppc is Big Endian architecture, no CERNLIB in modern systems, IBM FORTRAN is not for free, 32 bit systems are dying out etc.

Andrii Verbytskyi `https://indico.cern.ch/event/588219`

Data volume challenge easy; software and computing environment challenge hard

Four pillars of reusable computational research

I. Input data

What is your input data?

- input files
- input parameters

II. Analysis code

Which code analyses it?

- user code
- software frameworks

III. Computing environment

What is your environment?

- operating system
- database calls

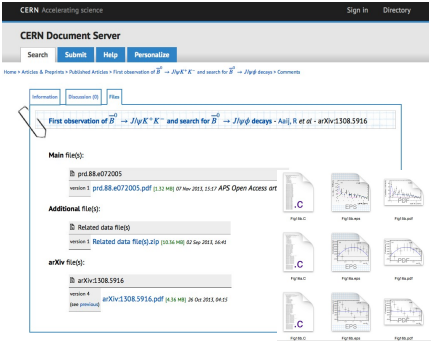
IV. Computational recipes

Which steps did you take?

- shell commands
- notebooks and workflows

I. Data: Publication-level data

LHCb collaboration `https://arxiv.org/abs/1308.5916`

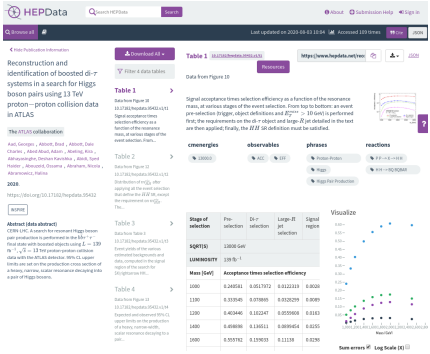


LHCb collaboration `https://cds.cern.ch/record/159565`

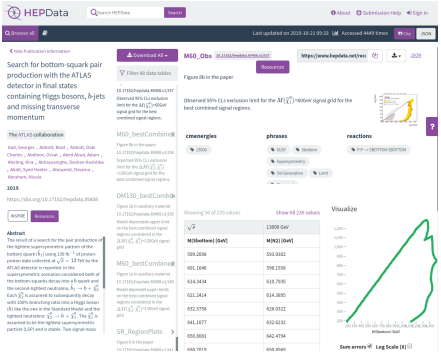
Experimental collaborations publishing papers on arXiv and INSPIRE

. . . and storing related data files in separate digital repositories

I. Data: Widening the publication-level data scope



ATLAS collaboration `https://www.hepdata.net/record/ins1809175`

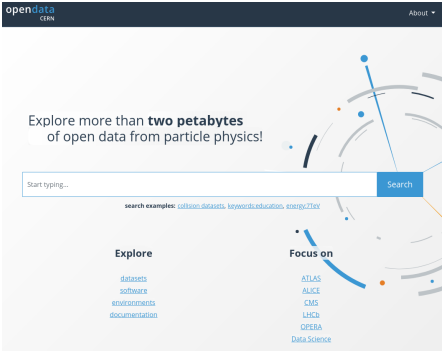


ATLAS collaboration `https://www.hepdata.net/record/ins1748602`

HEPData provides interactive interface to explore and download publication-level data behind plots and tables

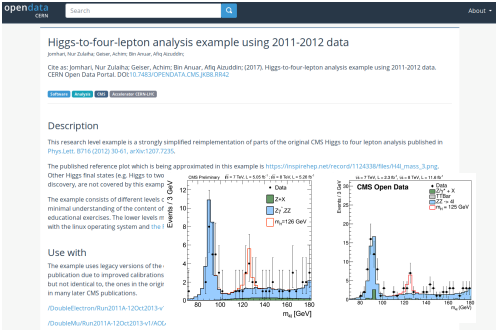
. . . now starting to handle more data types: likelihoods!

I. Data: Event-level open data for education and research



`http://opendata.cern.ch`

CERN Open Data portal contains more than two petabytes of research-grade raw, collision and simulated data

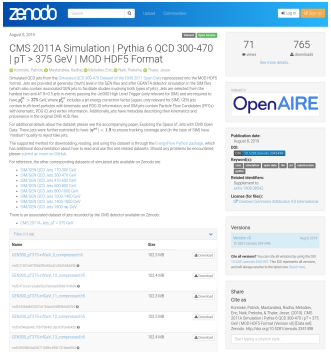


N. Z. Jomhari *et al* `http://opendata.cern.ch/record/5500`

Data is released together with code examples and Virtual Machine images enabling to run realistic physics analysis examples

I. Data: Machine Learning community

Di Florio *et al* `http://opendata.cern.ch/record/12320`



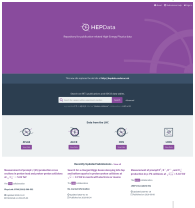
Komiske *et al* `https://zenodo.org/record/3341498`

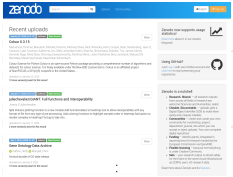
Machine-learning related datasets on the CERN Open Data portal

Machine-learning related datasets on Zenodo

A family of digital repositories

publication-level data



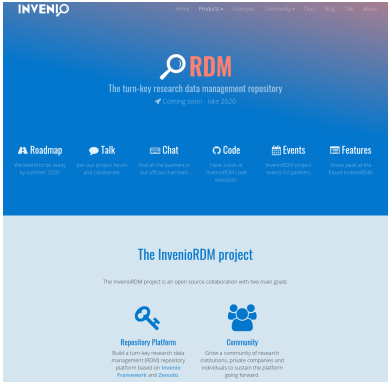


experimental collaborations

invidivual researchers

event-level data

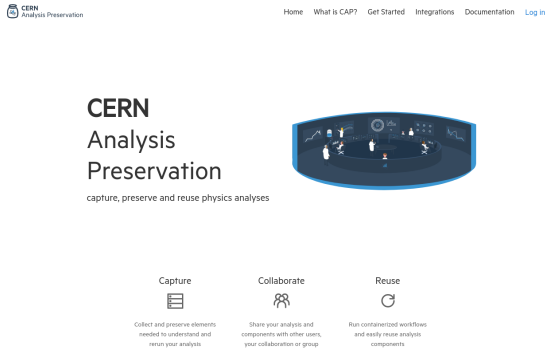
Invenio digital repository framework

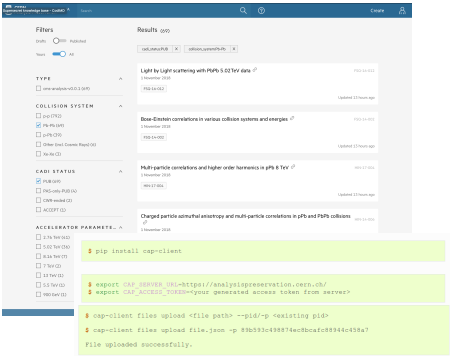


https://invenio-software.org/products/rdm

InvenioRDM project building a turn-key solution for Research Data Management

I. Data: Preserving “restricted” analysis knowledge

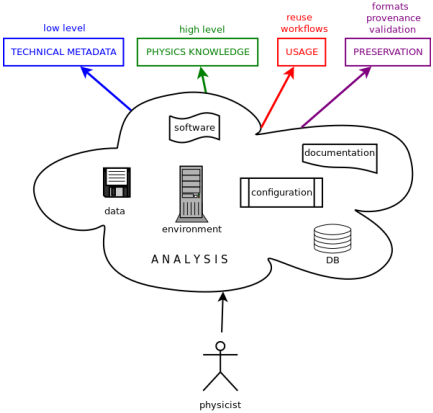


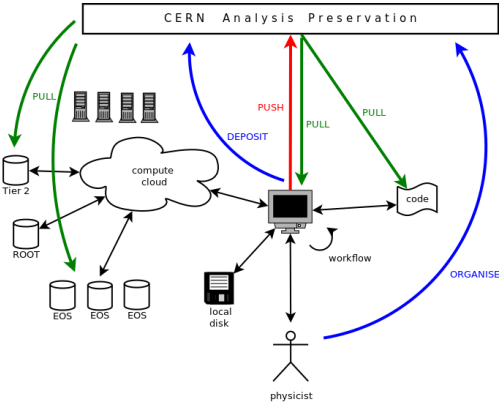


`https://analysispreservation.cern.ch`

CERN Analysis Preservation service for collaboration-restricted data.
Following the same FAIR principles (FAIR $\neq$ open!)

Structuring and capturing knowledge behind particle physics data analyses





Structuring knowledge behind research data analysis (custom JSON schemas)

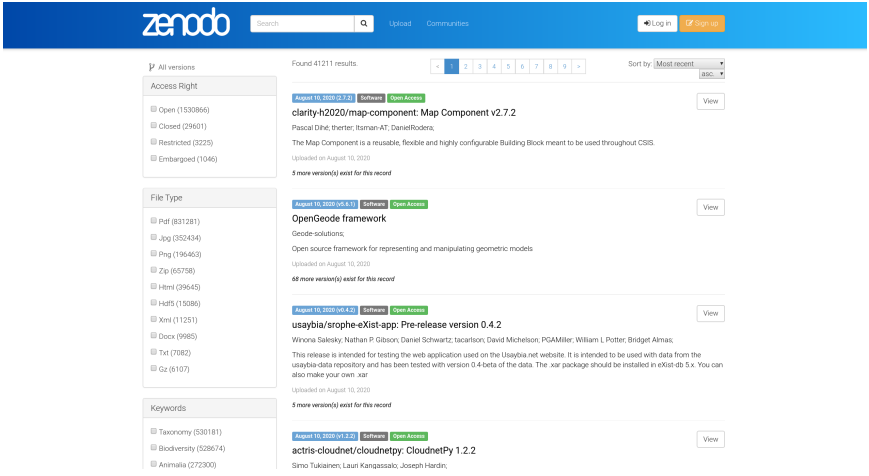
Taking consistent snapshot of analysis assets at a certain time (via “push” or “pull”)

II. Code: Preserving research software

https://guides.github.com/activities/citable-code

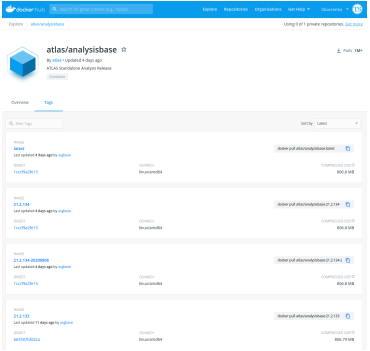
GitHub ↔ Zenodo bridge to automatically preserve releases

II. Code: Preserving research software



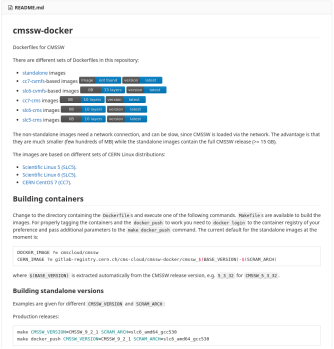
Zenodo minted more than 90,000 DOIs for software

III. Computing environment: containers



ATLAS collaboration

`https://hub.docker.com/r/atlas/analysisbase/tags`

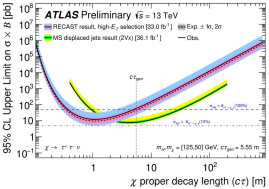


CMS collaboration

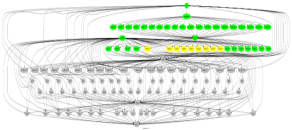
`https://gitlab.cern.ch/cms-cloud/cmssw-docker`

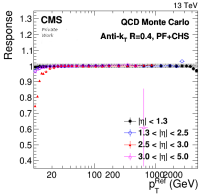
Docker and Singularity images help to encapsulate the computing environment

IV. Computational recipes: notebooks & workflows



ATLAS `http://cdsweb.cern.ch/record/2714064`



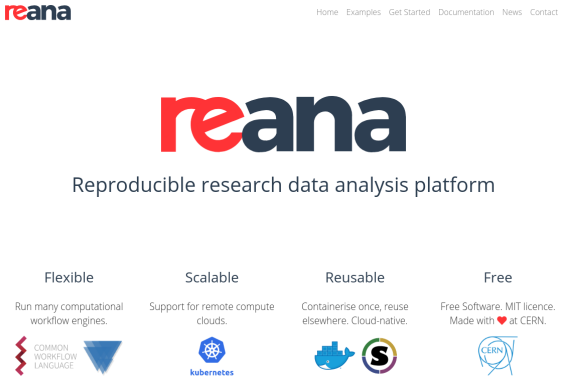


CMS `https://github.com/alintulu/reana-demo-JetMETAnalysis`

Data analysis example: ATLAS displaced jet search reinterpretation

Data production example: CMS jet energy resolution and corrections

REANA: A platform for reproducible analyses



http://www.reana.io

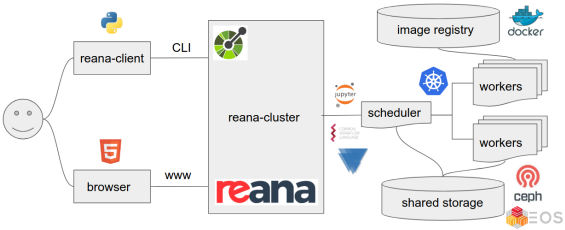
Install and run containerised workflows on Kubernetes, HTCondor, Slurm backends

REANA: Running containerised workflows on remote compute clouds

```yaml
version: 0.6.0
inputs:
  files:
    - skim.cxx
    - histograms.py
    - plot.py
workflow:
  type: serial
  specification:
    steps:
      - name: compiling
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - g++ -g -O3 -Wall -Wextra -Wpedantic -o skim skim.cxx `root-config --cflags --libs`
      - name: skimming
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - ./skim
      - name: histogramming
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - python histograms.py
      - name: plotting
        environment: reanahub/reana-env-root6:6.18.04
        commands:
          - python plot.py
```

https://github.com/cms-opendata-analyses

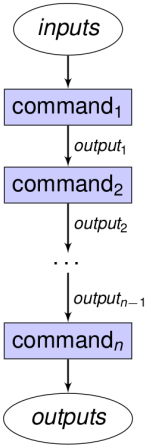
Structure data analysis by means of a `reana.yaml` file



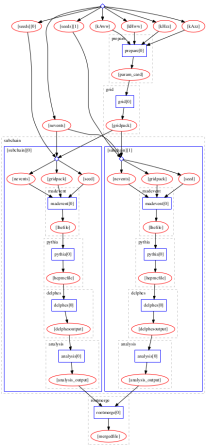
https://doi.org/10.1051/epjconf/201921406034

Use command-line and web interfaces to launch analysis on remote compute clusters

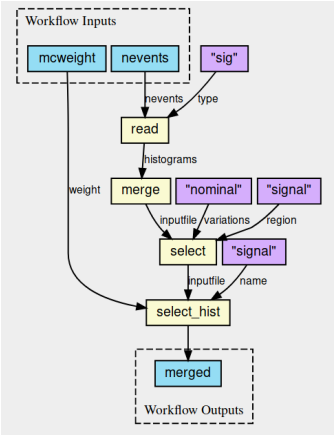
Computational workflows



Serial



Yadage



CWL

Fostering FAIR science

CERN Open Data

CERN Analysis Preservation

REANA Reusable Analyses

"Capturing and sharing actionable data and knowledge behind particle physics data analyses in order to facilitate future data reuse"

CERN IT J. B. Gonzalez Lopez, A. Lintuluoto, D. Rodríguez, P. Shandilya, T. Šimko, M. Vidal García · **CERN SIS** P. Fokianos, I. Koutsakis, A. Lavasa, K. Naim, A. Papadopoulos, S. van de Sandt · **ALICE** Y. Foka, M. Gheata, C. Grigoras, M. Zimmermann · **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** H. Bittencourt, A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, S. Roiser, A. Trisovic · **OPERA** G. De Lellis, S. Dmitrievsky, G. Galati, A. Ustyuzhanin · **HepData** A. Clarke, E. Maguire, G. Watt · **Invenio** L. H. Nielsen, P. Panero · **Zenodo** F. Decourcelle, A. Ioannidis, G. Lignos · **CERN CernVM** J. Blomer · **CERN EOS** L. Mascetti, H. Rousseau · **CERN Kubernetes** R. Rocha · **CERN OpenShift** A. Lossent