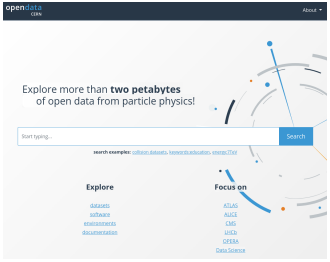
CERN IT and HEPData

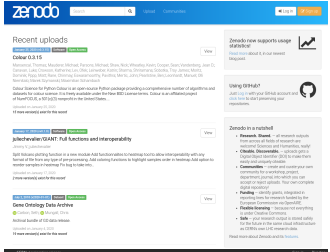
Tibor Šimko

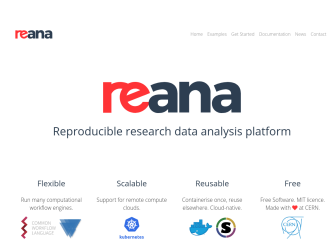
CERN

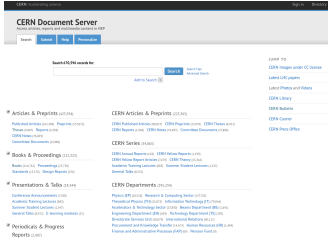
HEPData Advisory Board Meeting, Durham, UK, 28 January 2020

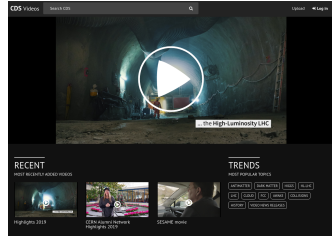
Research data repository related activities in CERN IT









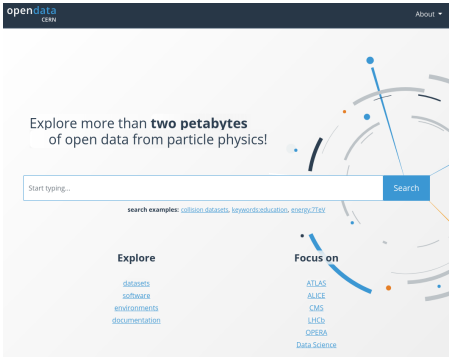


CERN Open Data

- *event-level data for education and research*
- collision, simulated, and derived datasets
- virtual machines and container images
- software tools and analysis examples
- configuration files and documentation
- over 7K records, 800K files, 2P bytes

HEPData

- classic content: *publication-level data*
- natural distinction and collaboration
- recent evolution: likelihoods; additional data?



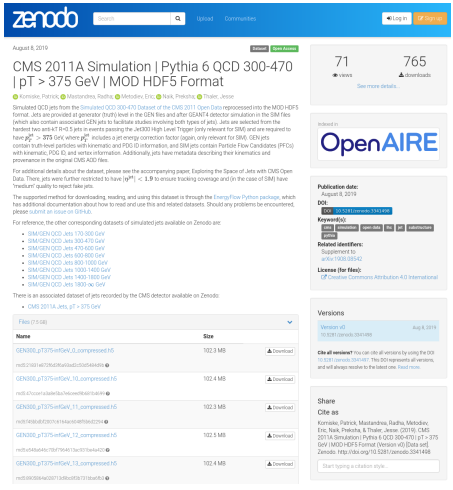
http://opendata.cern.ch

Zenodo

- long-tail of science; all scientific disciplines
- over 1.5M of data, software, document records
- over 70K DOIs for software
- used by individuals and self-organised communities

HEPData

- cross-experiment communities such as Machine Learning?
- some ML data on CERN Open Data
- some ML data on Zenodo

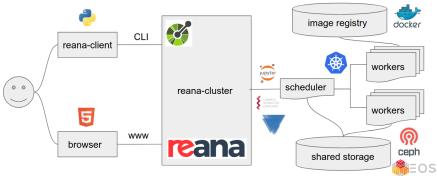


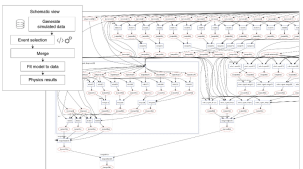
REANA reproducible analysis platform

- run containerised analysis pipelines on remote compute clouds
- data + code + environment + workflows = = reproducible science
- FAIR data reuse

HEPData

- push workfow assets to CERN Analysis Preservation, Zenodo . . . HEPData
- pull information from HEPData
- integration in automated data analysis and reuse workflows?



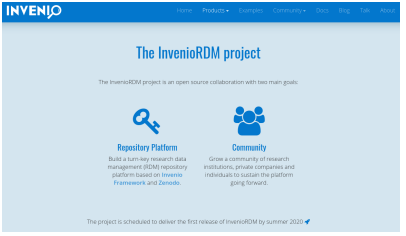


Invenio digital repository framework

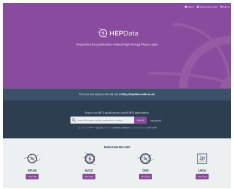
- Mature digital repository framework
- used by CERN Analysis Preservation, CERN Document Server, CERN Open Data, HEPData, INSPIRE, Zenodo ...
- started generalised Invenio RDM project (2019-)

HEPData

- using Invenio lightly (`records` module)
- interest in certain modules?
- digital repository framework needs?

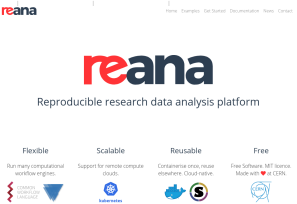


Conclusions: technology-synergies



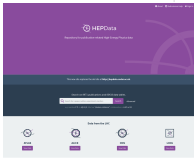
pull

push



repository framework

Conclusions: content-synergies



publication-level data

experimental collaborations ⟷ invidivual researchers

event-level data