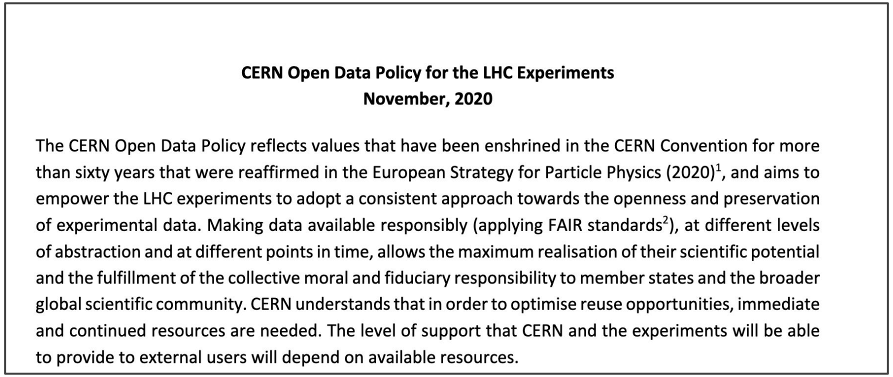
Streamlining opendata policies in Rucio data management platform

Hugo Gonzalez Labrador
Storage Engineer
CERN IT

7th Rucio Workshop, 02/10/2024

Setting the scene

LHC experiments at CERN publish a significant amount of their collected and derived data as open data products, as mandated by the CERN Open Data Policy for CERN experiments.



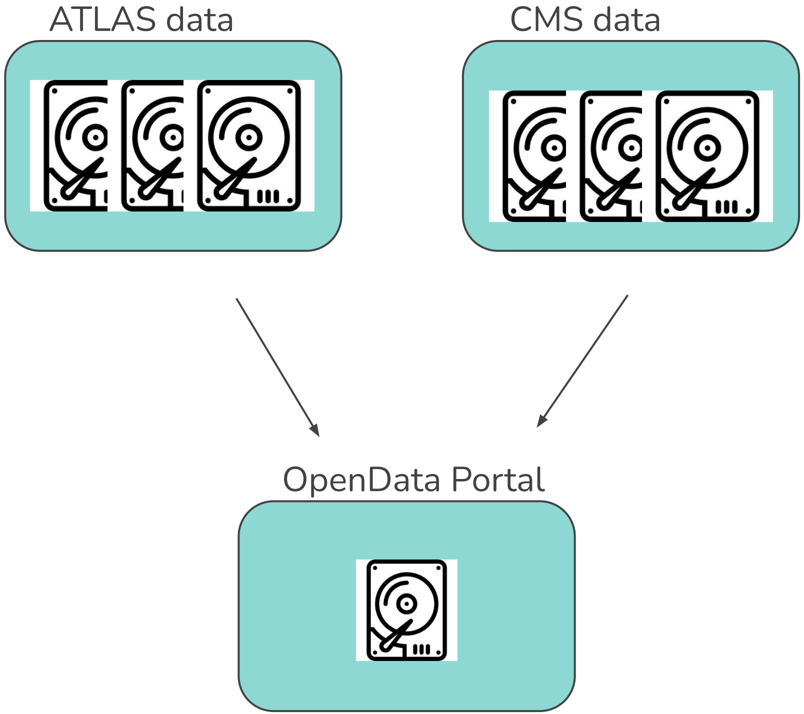
CERN Open Data Policy for the LHC Experiments
November, 2020

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)[1], and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAIR standards[2]), at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community. CERN understands that in order to optimise reuse opportunities, immediate and continued resources are needed. The level of support that CERN and the experiments will be able to provide to external users will depend on available resources.

https://cds.cern.ch/record/2745133/files/CERN-OPEN-2020-013.pdf

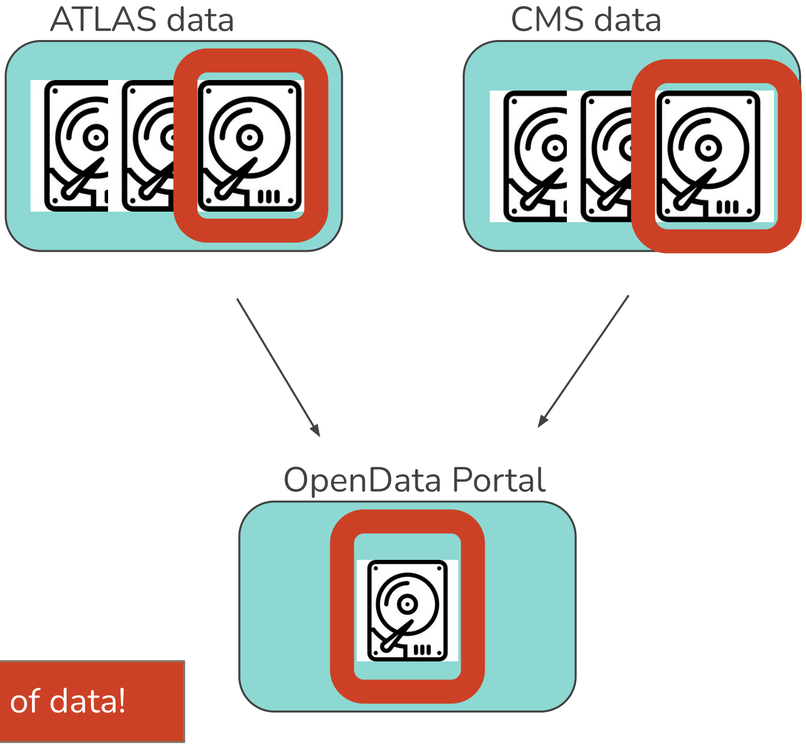
Example at CERN

The current amount of **open data released** by these two experiments is **10 petabytes**,, accounting for more than **85% of the overall open data available through the CERN open data portal.**

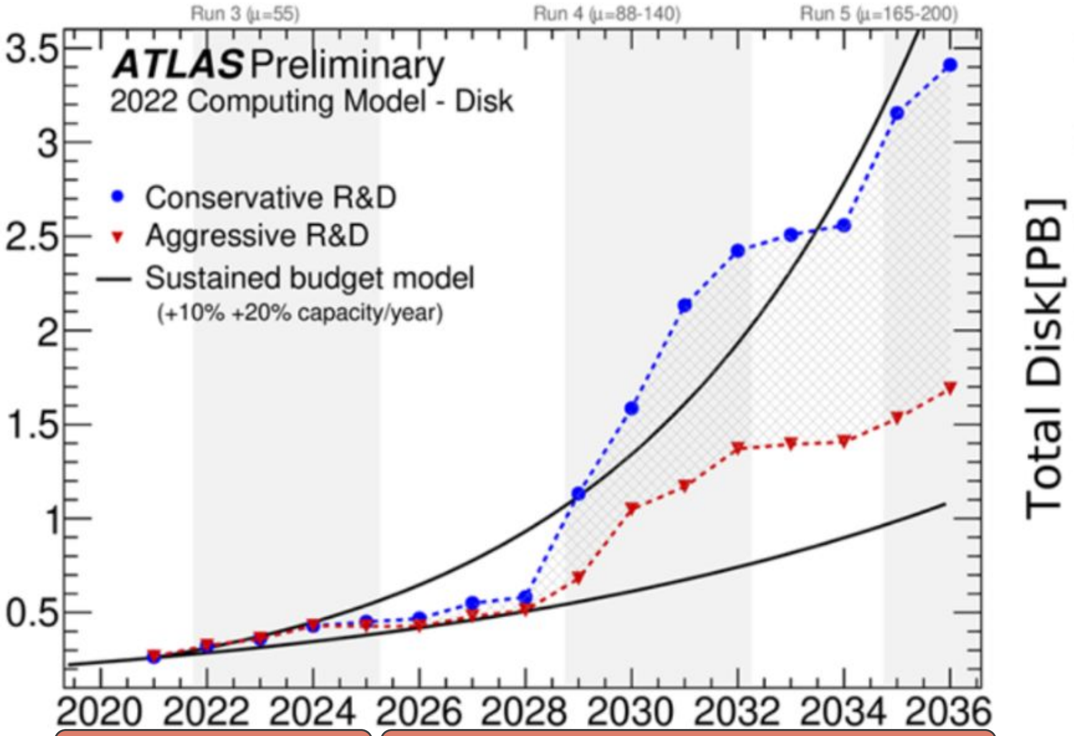


Example at CERN

The current amount of **open data released** by these two experiments is **10 petabytes**,, accounting for more than **85% of the overall open data available through the CERN open data portal.**



Potential duplication of data!



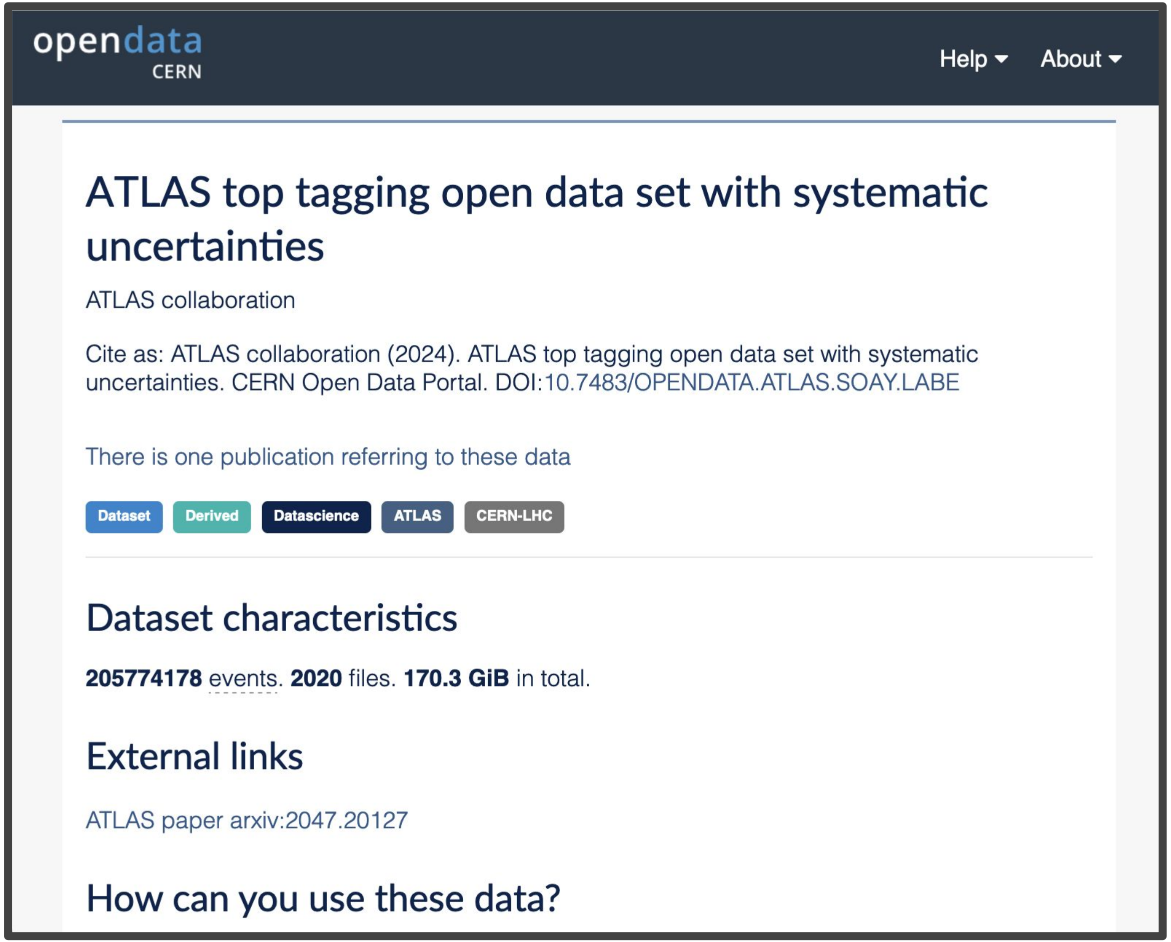
Backlog

Future backlog

2012-2020

Some open data already
released in this period

Can we avoid copying data to 3rd party systems to avoid duplication and still have FAIR principles on data to make it open?



ATLAS top tagging open data set with systematic uncertainties

ATLAS collaboration

Cite as: ATLAS collaboration (2024). ATLAS top tagging open data set with systematic uncertainties. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.SOAY.LABE

There is one publication referring to these data

Dataset characteristics

205774178 events. **2020** files. **170.3 GiB** in total.

External links

ATLAS paper arxiv:2047.20127

How can you use these data?

Project proposal for OSCARS EU call

Project title: Streamlining opendata policies in R

Principal investigator and team:

- Hugo Gonzalez, CERN IT (PI)
- Martin Barisits, CERN EP
- DevOps Engineer to hire

Start/End Date: 01.01.2025 – 31.12.2026

Funding: 250,000 euros

APPROVED
★ ★ ★
APPROVED
★ ★ ★
APPROVED

Future impact of the project

> Reduce storage costs by avoiding copy of data from experiment’s storage to 3rd party systems

> Facilitate ingestion of data and metadata to open data portals (ex: CERN opendata, data.europa.eu)

> Rucio already stores ***some*** metadata from experiments catalogues

> Data does not leave experiments data mgmt systems, knowledge and pledges

Future impact of the project

Rucio is used by many other scientific collaborations, many represented in this workshop: DUNE, AMS, Belle II, ICARUS, LIGO/VIRGO, CTA, MAGIC, Rubin LSST (to name a few) **that will also probably need to to release open data**

Current practices are manual and costly (ex, sharing open data via Dropbox link, no FAIR principles, etc…)

This proposal initiated from roadmap of ESCAPE collaboration on best data management practices, the **benefits are expected to be exploitable by multiple Rucio communities**

Innovation points

Paradigm shift: linking data vs copying data

Open data can be anywhere on the grid without changes to underlying storage systems

Lower the barrier to publish open data

Direct integration of open data inside experiment’s DDM systems

Extending the data lifecycle for data products

Policies can be changed at anytime for best resource usage and carbon footprint (move copies from green DC to greener DC)

Innovation points

- Bring open data closer to end-users
- **Provides embargo policies for access to the data**
- Rucio SDN connects users with closest geographically storage systems
- Facilitates FAIR principles towards open data repositories
- **Enables cross-experiment data sharing without incurring copying data**
  - Useful for cross-experiment validation, for example, re-using MC simulations
- Increase availability of open data from single EFRIs: “open data will have 3 copies on tape in 3 different continents”

Objectives

OBJ1: **Research and implement capabilities in Rucio to support a new class of data (open data)** to ensure policies for data products marked as "opendata" are honoured (copies on custodial storage, copies on a specific storage systems, number of copies in total, etc.)

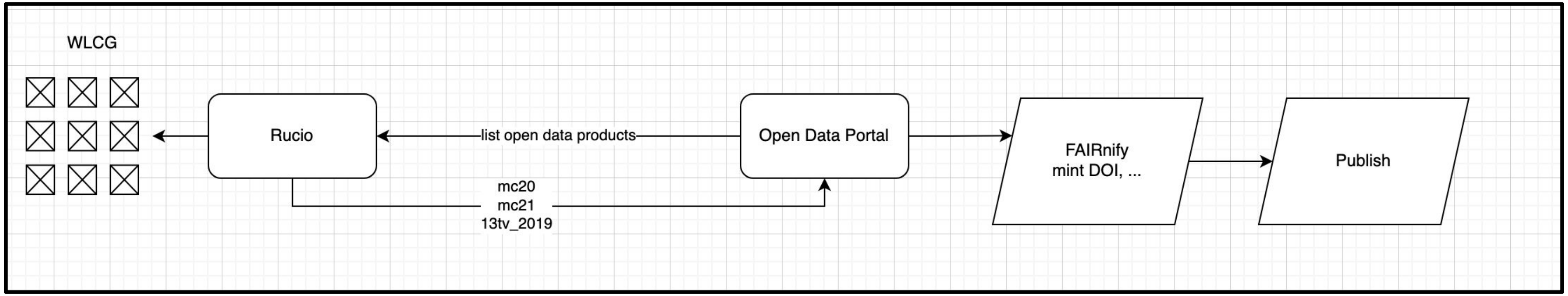
OBJ2: Develop novel Rucio open data interface, open data catalogues and platforms can request a list of declared open data in Rucio. Additionally, public users, through open data catalogues/platforms, can request access to Rucio-declared open data. Access is given specifically to the requested open data files via capability restricted oAuth2 tokens. Thus, it is ensured that only open data files can be read publicly directly from the experiments storage areas.

Objectives

OBJ3: Analyze current long-term data preservation workflows (selection, curation, preservation and archiving) and provide guidance on how to achieve similar requirements natively with Rucio policies and facilitate data ingestion from open data platforms.

OBJ4: Demonstrate technical viability of cross-system data sharing across science domains, useful for multiwavelength research. Demonstrator within ATLAS and CMS experiments for MC production data sharing.

Challenges



- Not all metadata needed to label data as open is available in Rucio
  - Usually this information lives in external catalogues: AMI (ATLAS), DBS (CMS) and we've seen this is a common practice across other experiments in the Rucio community
  - How to solve this?
- The Open Data portal will need to have exclusive access to the data tagged as "open data", i.e users and admins should not be able to mess with it (modification requires a new DOI and publishing step)
- Metadata filtering is in early stages, improvements ongoing by Dimitris, see his talk

Sounds interesting?
Similar challenges in your organization?

Get in touch:
hugo.gonzalez.labrador@cern.ch
martin.barisits@cern.ch