CERN Open Data: Policy to Implementation

J. Boyd obo the CERN OD WG

DPHEP workshop @ CERN

3/10/2024

Many related talks at this workshop:
ALICE: David Chinellato
ATLAS: Zach Marshall
CMS: Julie Hogan
LHCb: Dillon Fitzgerald
LEP: Jacopo Fanini , Dietrich Liko, Matthias Schroeder, Ulrich Schwickerath
OD portal: Pablo Saiz

Introduction

Some brief history:

Open Data Policy Working Group for the LHC expts

- Iniated by Eckhard Elsen (DRC at the time) in Feb 2020
- Working Group consists of 2 representatives each from the large LHC experiments (1 from TOTEM), and a representative from IT and SIS
- WG chaired by Jamie Boyd
- Mandated to draft a policy document relevant to the LHC experiments Open Data

- Policy endorsed by large LHC experiments by end of 2020
- Since then Joachim Mnich (DRC) has asked me to work on expanding the set of experiments that sign up to the policy
  - First "small" LHC experiments (TOTEM, LHCf, MoEDAL,FASER, SND@LHC)
    - Now all of these experiments have endorsed the policy!
  - Second, non-LHC experiments
    - Discussions with these experiments ongoing….

The ODWG is now a standing WG, with once yearly meetings to discuss the status of the policy implementation.

Policy Strategy

- The policy has been broken down into the 4 levels of data as defined in the DHEP study on data preservation:
  - Level 1 – Scientific publications, and associated additional data
  - Level 2 – Data useful for Education and Outreach
  - Level 3 – Reconstruction level data useful for general physics analysis
  - Level 4 – RAW data
- All large-LHC experiments already released data for L1 and L2 in broadly similar ways, and all agree that L4 is not practically useful
  - The WG discussion therefore strongly focussed on the policy for L3 data
- Decided to write a short general policy document (to be made public) outlining the general principles, and in addition an internal document outlining the implementation of the policy by each experiment
  - Implementation strategy for each experiment cannot be changed without discussing in the WG

Policy document : https://cds.cern.ch/record/2745133

Implementation document : https://cds.cern.ch/record/2745081

Public Policy Document

- 2 page document:
- Introduction:
  - Motivation and scope
  - Outline DHEP levels of data
- L1: Published results policy:
  - Publish results in Open Access journals
  - Provide additional information through HepData / RIVET etc..
  - No change in current policy
- L2: Outreach and Education
  - Provide rich data samples tailored for outreach/education (easy to use)
- L3: Reconstructed data
  - See next slides…
  - Real change to Collaborations policy
- L4: Raw Data
  - Not useful for external people

CERN Open Data Policy for the LHC Experiments
September 24, 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.

This policy relates to the data collected by the LHC experiments, for the main physics programme of the LHC — high-energy proton–proton and heavy-ion collision data. The foreseen use cases of the Open Data include reinterpretation and reanalysis of physics results, education and outreach, data analysis for technical and algorithmic developments and physics research. The Open Data will be released through the CERN Open Data Portal which will be supported by CERN for the lifetime of the data. The data will be tailored to the different uses, and will be made available in formats defined by each experiment that afford a range of opportunities for long-term use, reuse and preservation. In general, four levels of complexity of HEP data have been identified by the Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group[3], which serve varying audiences and imply a diversity of openness solutions and practices.

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,[4] with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

Outreach and Education (Level 2) Policy: For the purposes of education and outreach, dedicated subsets of data are used, selected and formatted to provide rich samples to maximise their educational impact, and to facilitate the easy use of the data. These data are released with a schedule and scope determined by each experiment. The data are provided in simplified, portable and self-contained formats suitable for educational and public understanding purposes; but are not intended nor adequate for the publication of scientific results. Lightweight environments to allow the easy

also be provided. CERN experiments will make data of such high level of
e through the CERN Open Data Portal.[5]

Policy: The LHC experiments will release calibrated reconstructed data for algorithmic, performance and physics studies. The release of these provenance metadata, and by a concurrent release of appropriate ware, reproducible example analysis workflows, and documentation. nts that are compatible with the data and software will be made ovided will be sufficient to allow high-quality analysis of the data olication of the main correction factors and corresponding systematic ations, detector reconstruction and identification. A limited level of vel 3 Open Data will be provided on a best-effort basis by the

periodically, following an appropriate latency period to allow thorough ne reconstruction and calibrations, as well as to allow time for the data by the collaboration. The size of the released datasets will be amount of data collected of similar type, with the aim to commence s of the conclusion of the run period. Data may be withheld by an analyses ongoing. Full datasets will be made available at the close of the

The data will be released from the CERN Open Data Portal under the Creative Commons CC0 waiver, and will be identified with persistent data identifiers, and the data must be cited through these identifiers. Similarly, appropriate acknowledgements of the experiment(s) should be included in publications released using such data, and the publications made clearly distinguishable from those released by the collaboration. Any scientific claims in such publications are the responsibility of their authors and not of the experiments. It is expected that scientific results released using Open Data follow best scientific practices. The experiments may impose rules related to the use of the data by members of their respective collaborations.

External authors should be aware that they will not have access to the vast amount of tacit knowledge built up within the LHC collaborations over the decades of design, construction and operation of the experimental apparatus. To allow external scientists to fully benefit from all the data, knowledge and tools, the collaborations may offer appropriate association programmes.

Raw Data (Level 4) Policy: It is not practically possible to make the full raw data-set from the LHC experiments usable in a meaningful way outside the collaborations. This is due to the complexity of the data, metadata and software, the required knowledge of the detector itself and the methods of reconstruction, the extensive computing resources necessary and the access issues for the enormous volume of data stored in archival media. It should be noted that, for these reasons, general direct access to the raw data is not even available to individuals within the collaboration, and that instead the production of reconstructed data (i.e. Level-3 data) is performed centrally. Access to representative subsets of raw data—useful for example for studies in the machine learning domain and beyond—can be released together with Level-3 formats, at the discretion of each experiment.

[1] European Strategy Group (2020), '2020 Update of the European Strategy for Particle Physics'.
[2] FAIR Guiding Principles for scientific data management and stewardship. Available at: https://www.go-fair.org/fair-principles/.
[3] Data management plans are defined by the LHC experiments to address the long-term preservation of internal data products. See: Akopov et al., Status report of the DPHEP Study Group: Towards a global effort for sustainable data preservation in high energy physics. arXiv preprint arXiv:1205.4667 (2012).
[4] Repository for publication-related High-Energy Physics data: http://www.hepdata.net.
[5] CERN Open Data portal: http://opendata.cern.ch.

L3 data – discussion (1)

- Any rules relating to publically releasing L3 data need to be approved by each experiments Collaboration Board
- Generally tried to find a good balance between:

1. Making data openly available

And

2. Protecting the collaborations:
   - Avoiding collaboration members publishing with open data rather than a Collaboration paper
   - Having to deal with wild claims made by external analysts with their data
   - Not taking too much resources (both human and computing) from the Collaboration

- After much discussion converged on limiting of the amount of data released after a given time as a tool to make it unattractive for collaboration members to publish with Open Data
  - Here a common approach across experiments is needed since e.g. ATLAS rules could effect a CMS collaborator's behaviour in this regard
  - Exact fraction of data released after what latency experiment specific, but general principle common across experiments

L3 data – discussion (2)

- Summary of each experiments Latency shown below:

	ALICE	ATLAS	CMS	LHCb
Fraction of data released in: 5 yrs (6 yrs for CMS)	10%	25% (but limiting to <20% of the total data at that time)	50% (but limiting to <20% of the total data at that time)	50%
Fraction of data released in: 10 yrs	50%	50% (but limiting to <20% of the total data at that time)	100% (but limiting to <20% of the total data at that time)	100%
End-of-Collaboration	100%	100%	100%	100%

L3 data – discussion (3)

- We should only release L3 Open Data if it can be used for high quality science
  - Release data and simulated samples
  - Release data with best calibrations available
  - Release sufficient information to allow main systematic uncertainties to be applied
  - Release analysis s/w to allow efficient analysis of data
- Data expected to be useable not only for particle physics, but also computing, algorithmic development, big data studies etc..
- Publications using Open Data should:
  - Have appropriate acknowledgements
  - Be clearly identifiable from Collaboration papers
  - Follow best scientific practices
- All experiments to use CERN Open Data Portal to house the released data

Implementation: Data Volumes

- Implementation document showed the expected media resource needs:

	ALICE (TB)	ATLAS (TB)	CMS (TB)	LHCb (TB)	SUM (TB)	Cumulative (PB)
2022	15	0	1089	200	1304	1.3
2023	20	150	872	4600	5642	6.9
2024	105	0	1436	0	1541	8.5
2025	105	0	1768	0	1873	10.4
Total	**245**	**150**	**5165**	**4800**	**10360**	

Table updated since, but broadly consistent with previous estimates.

- Currently ~5.5PB of LHC OD stored in the portal
- CERN IT kindly agreed to provide these resources for first 5 years (until end of 2025) - then to be re-discussed
- IT position to develop integrated tape back-end (save costs on storage resources in the long-term)
  - Work ongoing, hopefully to be deployed next year

Level-3 data so far released

ATLAS releases 65 TB of open data for research — 7/24
Explore over 75 billion LHC collision events — from home
News ATLAS

CMS releases 13 TeV proton collision data from 2016 — 4/24
CMS releases 13 TeV proton collision data from 2016
News CMS

LHCb releases the entire Run I dataset — 12/23
Today the LHCb collaboration completes the release of the data collected throughout the Run I of the Large Hadron Collider at CERN.
News LHCb

CMS completes Run-1 heavy ion open data collection — 9/23
New release of simulations, proton-lead collision data, and proton reference data.
News CMS

LHCb releases first set of data to the public — 12/22
The LHCb collaboration has released data from Run 1 of the LHC to the public for the first time, allowing research to be conducted by anyone in the world.
News LHCb

CMS completes the release of its entire Run-1 proton-proton data — 12/21
All proton-proton data collected by the CMS experiment during LHC Run-1 (2010-2012) are now available through the CERN Open Data Portal.
News CMS

First CMS open data from LHC Run 2 released
As the experiments at the Large Hadron Collider (LHC) brace for the start of Run 3 of the accelerator's programme in 2022, the CMS collaboration has released a ne
research-quality open data recorded by the CMS detector in 2015, the first year of Run 2. The new datasets are now available on the CERN Open Data portal.
News CMS

CMS releases heavy-ion data from 2010 and 2011 — 12/20
CMS releases heavy-ion data from 2010 and 2011
News CMS

Level-3 data so far released

ATLAS releases 65 TB of open data for research — 7/24
Explore over 75 billion LHC collision events — from home
News ATLAS

CMS releases 13 TeV proton collision data from 2016 — 4/24
CMS releases 13 TeV proton collision data from 2016
News CMS

LHCb releases the entire Ru
Today the LHCb collaboration co
News LHCb

CMS completes Run-1 heav
New release of simulations, proto
News CMS

LHCb releases first set of da
The LHCb collaboration has relea
News LHCb

CMS completes the release
All proton-proton data collected b
News CMS

No ALICE data release on Open Data portal yet, but will happen VERY soon

A lot of work has been going in ALICE to prepare for this:

- Developed new format to release data in
  - Needed for efficient long-term release of data
- Converted planned datasets to new format
- Established a new software framework for analysing converted data
- Validation of new samples now done
- In final stages of metadata preparation and data transfer to portal, for:
  - pp, p-Pb and Pb-Pb Run 1 data samples

12/21

First CMS open data from LHC Run 2 released
As the experiments at the Large Hadron Collider (LHC) brace for the start of Run 3 of the accelerator's programme in 2022, the CMS collaboration has released a ne
research-quality open data recorded by the CMS detector in 2015, the first year of Run 2. The new datasets are now available on the CERN Open Data portal.
News CMS

12/20

CMS releases heavy-ion data from 2010 and 2011
CMS releases heavy-ion data from 2010 and 2011
News CMS

Level 3 Open Data: LHCb

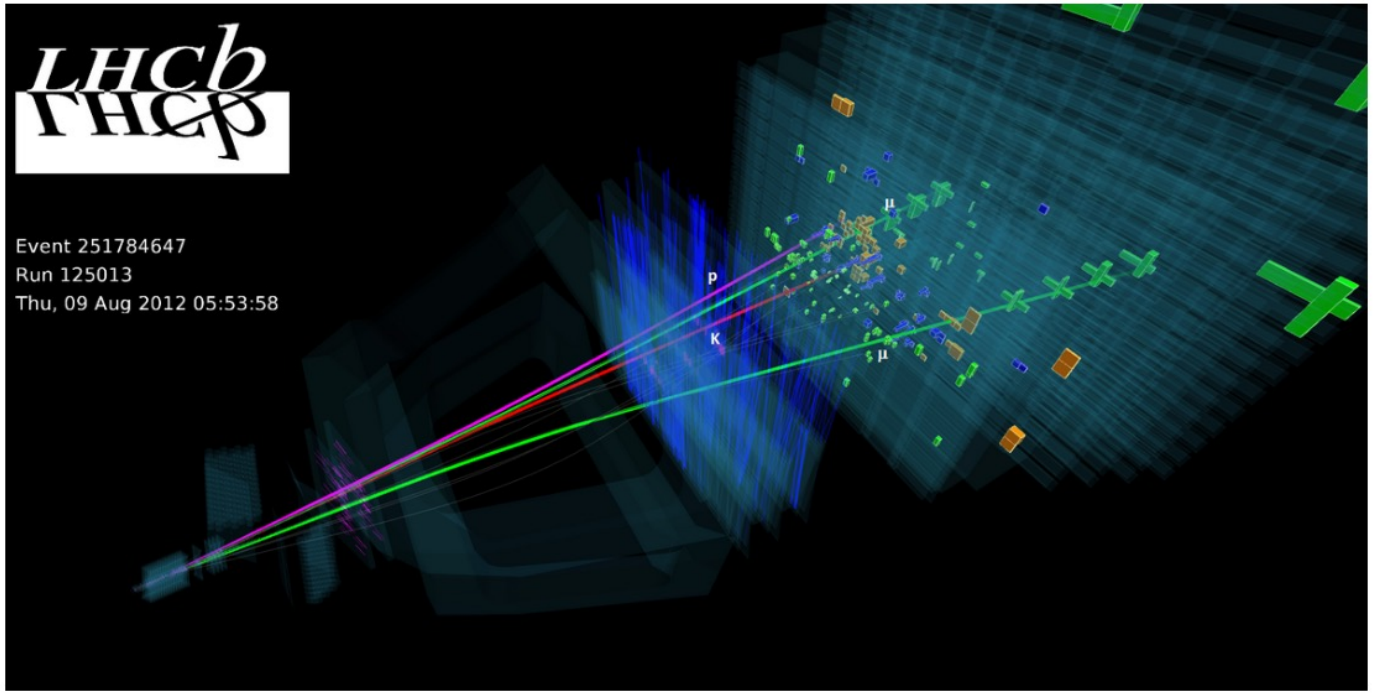
20/12/23

LHCb releases the entire Run I dataset

2023-12-20 by LHCb Collaboration

News

Today the LHCb collaboration completes the release of the data collected throughout the Run I of the Large Hadron Collider at CERN. The sample made available amounts to approximately 800 terabytes (TB) of data. These data, collected by the LHCb experiment in 2011 and 2012, contains information obtained from proton-proton collisions. The format made available provides pre-filtered data, suitable for a wide range of physics studies. The image below displays an event recorded during 2012.



Level 3 Open Data: CMS

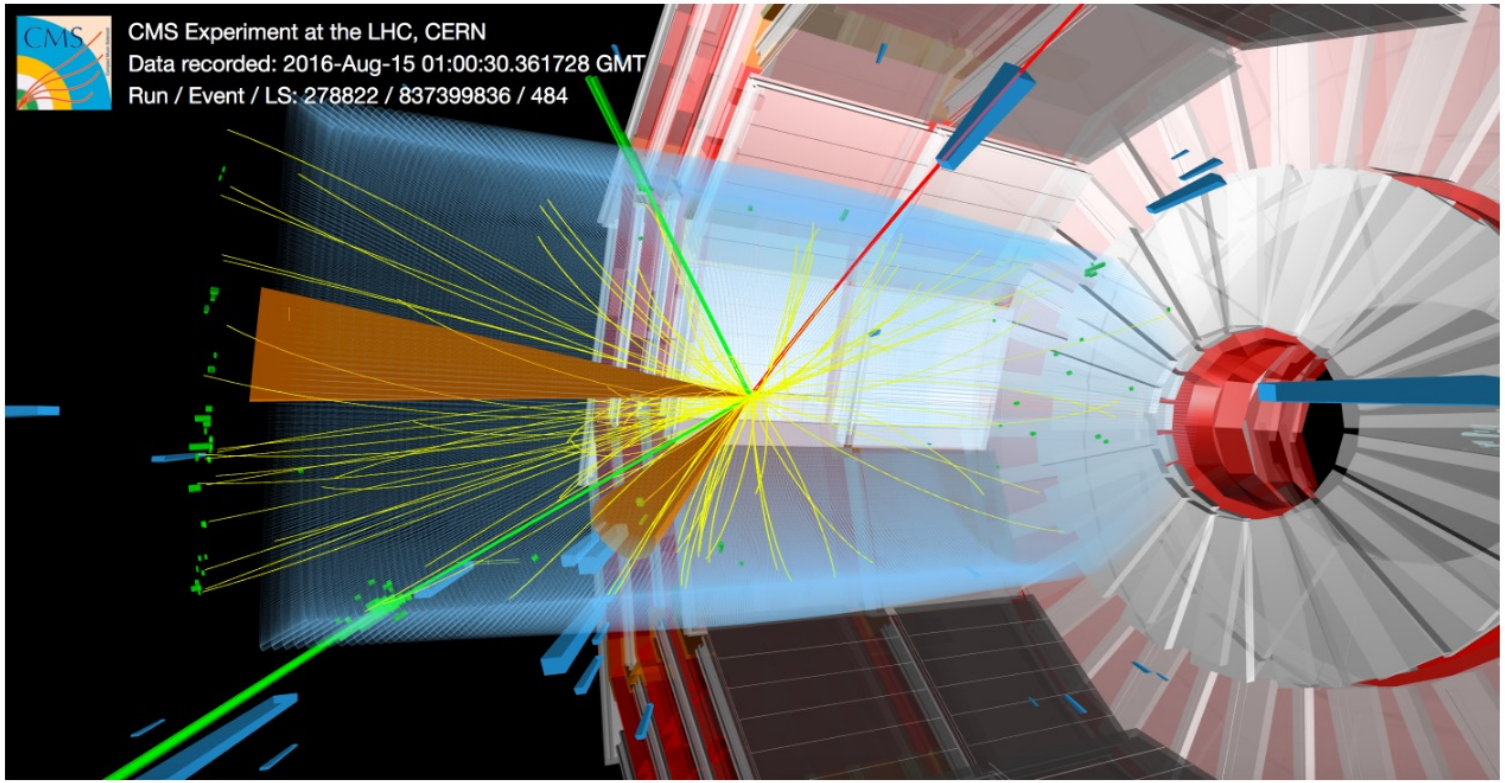
2/4/24

CMS releases 13 TeV proton collision data from 2016

2024-04-02 by CMS Collaboration

News

The CMS experiment at CERN is proud to announce the first release of 13 TeV proton-proton collision data collected in 2016. Over 70 TB of 13 TeV collision data and 830 TB of corresponding simulations are now accessible to the global scientific community and enthusiasts alike through the CERN Open Data Portal.



A candidate event in which a top quark is produced in association with a Z boson. Photo: CERN Document Server

Level 3 Open Data: ATLAS

1/7/24

ATLAS releases 65 TB of open data for research

2024-07-01 by ATLAS Collaboration

News

Explore over 7 billion LHC collision events – from home

The ATLAS Experiment at CERN has made two years' worth of scientific data available to the public for research purposes. The data include recordings of proton–proton collisions from the Large Hadron Collider (LHC) at a collision energy of 13 TeV. This is the first time that ATLAS has released data on this scale, and it marks a significant milestone in terms of public access and utilisation of LHC data.

"Open access is a core value of CERN and the ATLAS Collaboration," says Andreas Hoecker, ATLAS Spokesperson. "Since its beginning, ATLAS has strived to make its results fully accessible and reusable through open access archives such as arXiv and HepData. ATLAS has routinely released open data for educational purposes. Now, we're taking it one step further — inviting everyone to explore the data that led to our discoveries."

Released under the Creative Commons CC0 waiver, ATLAS has made public all the data collected by the experiment during the 2015 and 2016 proton–proton operation of the LHC. This is approximately 65 TB of data, representing over 7 billion LHC collision events. In addition, ATLAS has released 2 billion events of simulated "Monte Carlo" data, which are essential for carrying out a physics analysis.

External researchers, in particular, are encouraged to explore the ATLAS open data. "Along with the data, we have provided comprehensive documentation on several of our analyses, guiding users through our process step-by-step," says Zach Marshall, ATLAS Computing co-Coordinator. "These guides provide first-hand experience of working on a real ATLAS result, allowing anyone to test our tools, and evaluate the systematic uncertainties associated with the result for themselves."

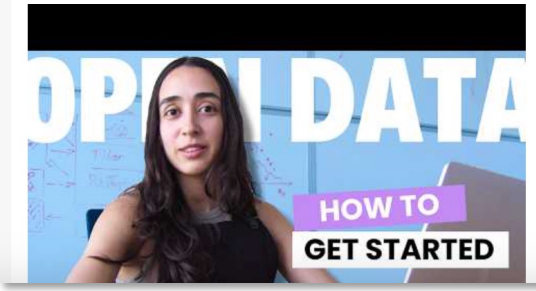
ATLAS traditionally collaborates with non-ATLAS scientists through short-term associations, granting them full access to ATLAS data, internal tools, and information. Through the open data, ATLAS researchers hope to further nurture this dialogue and collaboration. "In particular," adds Zach, "we'd like to encourage phenomenologists and also computer scientists to explore our datasets, instead of relying on mock-ups."

Today's release builds upon previous open data releases for educational use (in 2016 and 2020). "All of our open data releases are now available through the ATLAS open data website," says Dilia Portillo, ATLAS Outreach and Education co-Coordinator. "The website includes multi-level documentation, video tutorials and online tools aimed at the full-spectrum of users, from high school students to senior particle physics researchers. In addition, the software used to create the education-use open data has been released. This provides a seamless transition from the research open data to all the tutorials for outreach and education, including newly updated Higgs-boson discovery documentation. With a bit of time and dedication, you can go from being a relative novice to carrying out your own analysis."

The ATLAS open data website also serves as a hub for the community, which includes teachers, students, enthusiasts and, now, scientists. Anyone diving into the open data can also directly engage with ATLAS physicists, who are available to respond to user feedback and take suggestions.

This release marks the start of more to come, with ATLAS' first release of lead-lead-nuclei collision data data up next. The ATLAS Collaboration, along with the other main LHC experiment collaborations, has committed to making all of its data publicly accessible after a certain time. Openness is deeply ingrained in the culture of high-energy physics, enabling greater accessibility, reproducibility and better science.

Begin your journey with ATLAS open data by following the tutorial below.



Evolving model

- At the time of writing the implementation document in 2020, it became clear that the disk space needs for LHCb L3 Open Data will become prohibitive after Run 1
  - Estimated storage space 10PB (45PB) for Run 2 (3) if using the same model as Run 1
- LHCb have therefore developed a new strategy based on the NtupleWizard for access to LHCb OD
  - Allows a user to skim the full data (stored on LHCb resources) using LHCb CPU resources to produce an ntuple tailored for their use case
  - Smart idea to trade off CPU for storage
  - At advanced stage of development, to be deployed next year
  - More details in later talk by Dillon Fitzgerald
- Current ALICE model will also not scale for Run 3 dataset (would be 2PB/year)
  - Plan to publish skimmed and derived dataset for Run 3 (to be released in 2030+)

Open Data monitoring

- In the context of the CERN Open Science effort a discussion has started about what metrics should be monitored related to CERN Open Data and its usage
  - This is an ongoing discussion, but some current thoughts are shown below
- Some Key Performance Indicators will be made public in the Open Science report, while others will be monitored internally to understand the usage at a more detailed level
- Discussions are ongoing but can think of two sets of KPIs:
  - Related to data stored and accessed
    - Quantitative KPIs can be extracted from the OD portal
    - Dashboard under development in IT
  - Related to publications using CERN OD
    - Quantitative KPIs can be extracted from Inspire (assuming papers reference OD DOIs as they should)
  - This can miss the use of OD which does not lead to publications e.g. for education purposes
    - More difficult to come-up with quantitative KPIs for this, but maybe able to have more qualitative information related to this from e.g. surveying users

Expanding the effort – small-LHC expts

Expanding OD across CERN experiments

- After the succesful drafting of the OD policy for the large LHC experiments, I have been asked by the DRC to try to expand this policy to ultimatly cover all experiments at CERN (where possible)
- Suggested to start with the small LHC experiment:
  - Some of these experiments are quite different from the large LHC experiments, which may effect the need/use of OD (e.g. emulsion based, passive detectors, only operating during special running conditions)
  - Datasets expected to be small compared to large LHC experiments but experiments have less (human) resources for OD effort

All small-LHC experiments have now agreed to this:

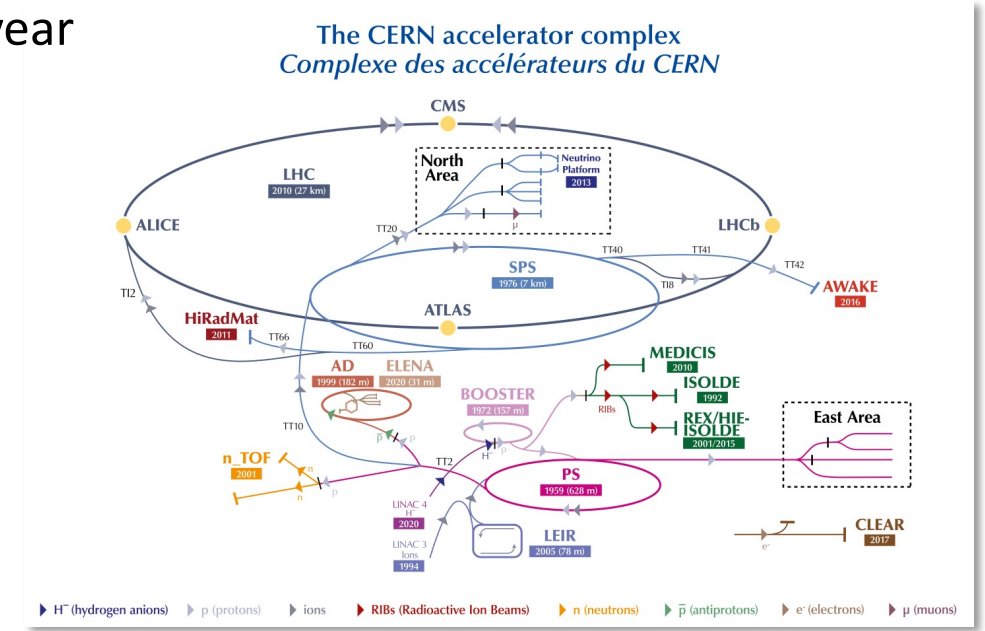
Experiment	Status	Comment
TOTEM	Policy endorsed	Operating since LHC start. Data taking coming to an end this year, OD way towards data preservation?
LHCf	Policy endorsed	Operating since LHC start. Only takes data in special runs. Last running during Run 3.
MoEDAL	Policy endorsed	In Run 1,2 only passive detectors. Installing electronic detectors during Run 3.
FASER	Policy endorsed	Only starting data taking in Run 3 (approved for duration of Run 3). Electronic detector, and (passive) emulsion detector data.
SND@LHC	Policy endorsed	

4

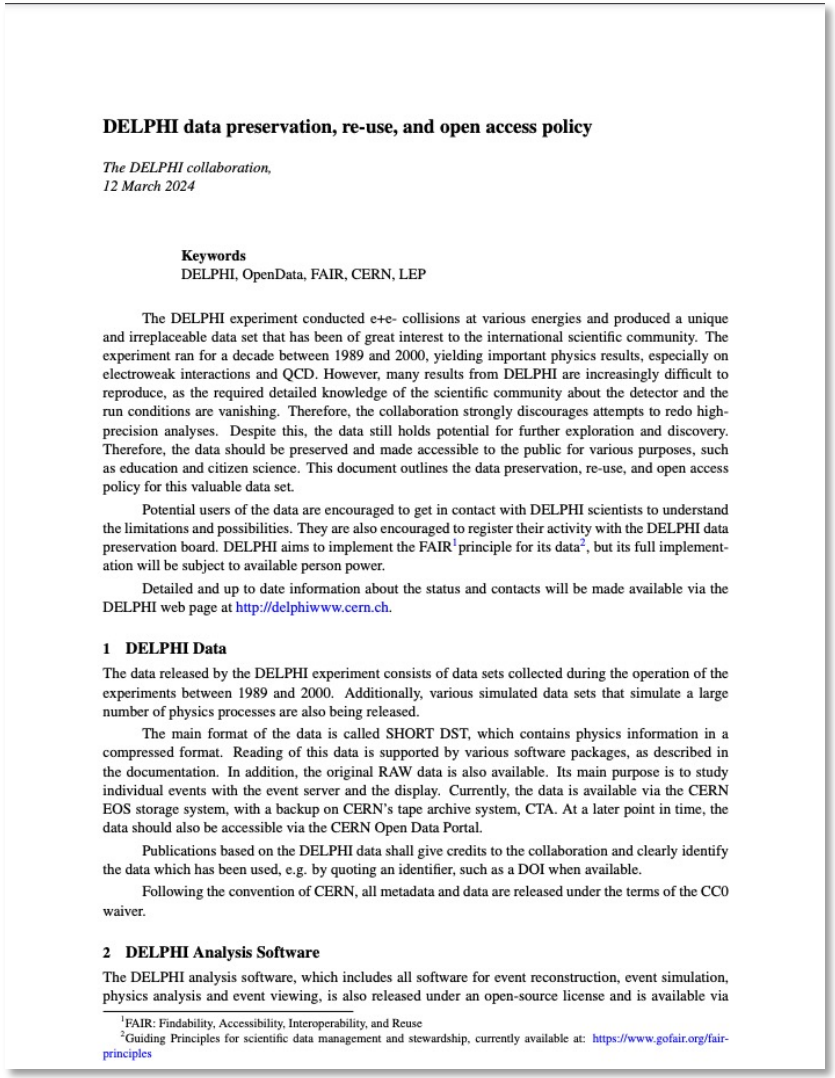
Given the 5 long latency before data needs to be released, these experiments do not need to put immediate effort into this yet

Expanding the effort – non-LHC expts

- Have started to discuss with non-LHC experiments, if they could sign-up to the existing LHC policy, or if they should develop a separate policy.
- Meeting organized with experiments Spokespersons on Spring this year
- For experiments associated with some CERN facilities the LHC policy, not well aligned with the procedures from the given physics community:
  - ISOLDE – have developed their own version of the policy:
    - https://isolde.cern/isolde-open-data-policy
  - nTOF – will likely develop their own policy, mostly based on existing procedures
  - AD experiments – to discuss if a common AD policy makes sense
- SPS based experiments mostly think LHC policy could work for them, but more discussion needed
  - Further discussion to be organized before the end of the year



Expanding the effort – LEP expts



DELPHI data preservation, re-use, and open access policy

The DELPHI collaboration,
12 March 2024

Keywords
DELPHI, OpenData, FAIR, CERN, LEP

The DELPHI experiment conducted e+e- collisions at various energies and produced a unique and irreplaceable data set that has been of great interest to the international scientific community. The experiment ran for a decade between 1989 and 2000, yielding important physics results, especially on electroweak interactions and QCD. However, many results from DELPHI are increasingly difficult to reproduce, as the required detailed knowledge of the scientific community about the detector and the run conditions are vanishing. Therefore, the collaboration strongly discourages attempts to redo high-precision analyses. Despite this, the data still holds potential for further exploration and discovery. Therefore, the data should be preserved and made accessible to the public for various purposes, such as education and citizen science. This document outlines the data preservation, re-use, and open access policy for this valuable data set.

Potential users of the data are encouraged to get in contact with DELPHI scientists to understand the limitations and possibilities. They are also encouraged to register their activity with the DELPHI data preservation board. DELPHI aims to implement the FAIR[1] principle for its data[2], but its full implementation will be subject to available person power.

Detailed and up to date information about the status and contacts will be made available via the DELPHI web page at http://delphiwww.cern.ch.

1 DELPHI Data

The data released by the DELPHI experiment consists of data sets collected during the operation of the experiments between 1989 and 2000. Additionally, various simulated data sets that simulate a large number of physics processes are also being released.

The main format of the data is called SHORT DST, which contains physics information in a compressed format. Reading of this data is supported by various software packages, as described in the documentation. In addition, the original RAW data is also available. Its main purpose is to study individual events with the event server and the display. Currently, the data is available via the CERN EOS storage system, with a backup on CERN's tape archive system, CTA. At a later point in time, the data should also be accessible via the CERN Open Data Portal.

Publications based on the DELPHI data shall give credits to the collaboration and clearly identify the data which has been used, e.g. by quoting an identifier, such as a DOI when available.

Following the convention of CERN, all metadata and data are released under the terms of the CC0 waiver.

2 DELPHI Analysis Software

The DELPHI analysis software, which includes all software for event reconstruction, event simulation, physics analysis and event viewing, is also released under an open-source license and is available via

[1] FAIR: Findability, Accessibility, Interoperability, and Reuse
[2] Guiding Principles for scientific data management and stewardship, currently available at: https://www.gofair.org/fair-principles

LEP data represents the highest energy e+e- collision data, very valuable – especially in the context of future e+e- machines (FCC-ee etc…)

ALEPH:

- Open Data available and used in some physics papers
  - (not on CERN OpenData portal though)

DELPHI:

- Recently: s/w ported to latest OS, released Open Data policy
- Data starting to be uploaded to portal

OPEL:

- Renewed effort started recently
- Recently : s/w ported to latest OS, policy being discussed in collaboration

L3:

- No known effort ongoing

https://delphi-www.web.cern.ch/delphi-www/delsec/finalrules/DELPHI_Data_preservation-8.pdf
https://dphep.web.cern.ch/experiment/aleph

Summary

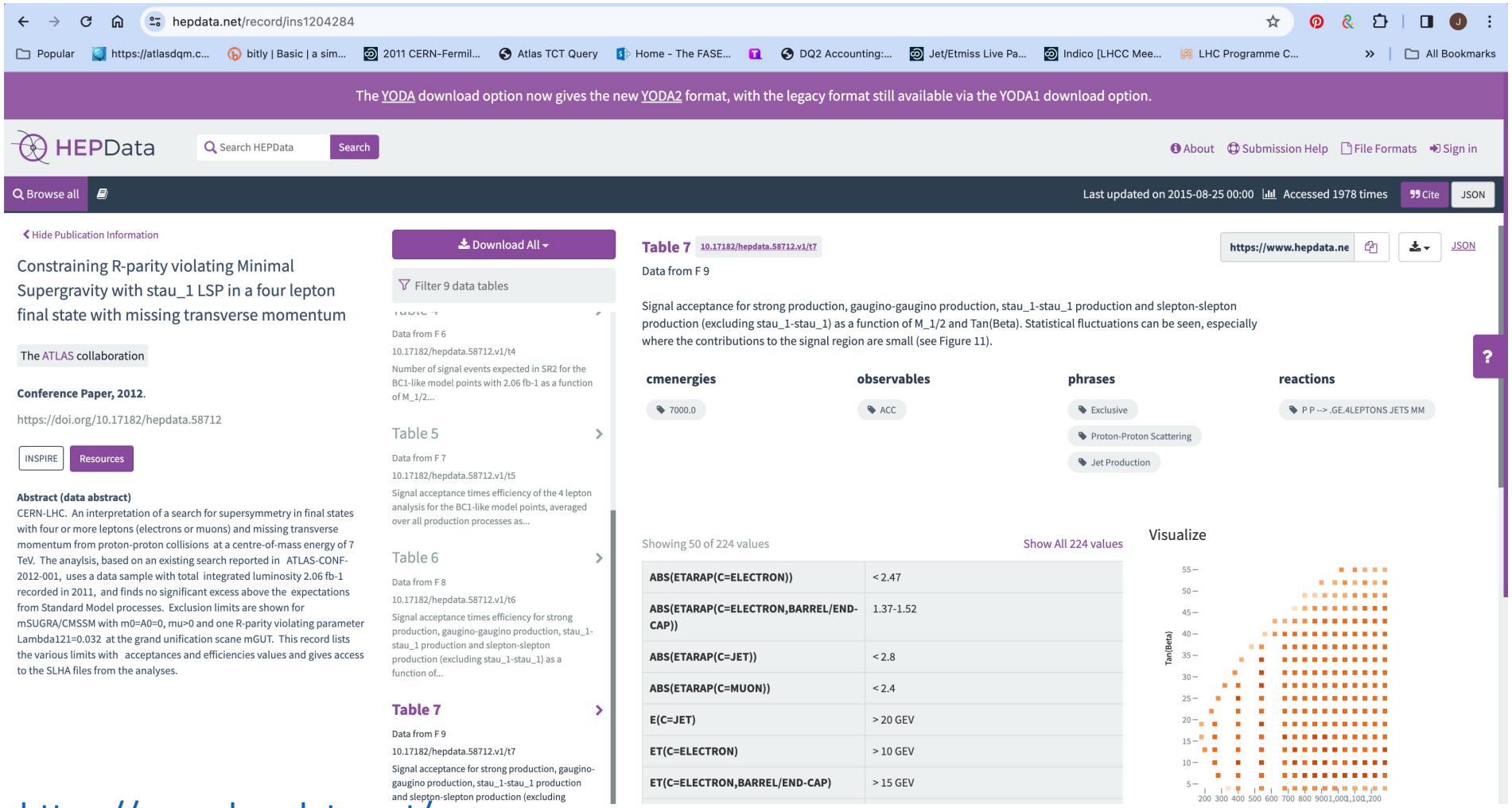
- LHC Open Data policy released at end of 2020, now entering implementation period
  - The ODWG standing WG to follow implementation of policy at a high level
    - Should expect some aspects to evolve as we see how this works in practise
  - All large LHC experiments have now (or are very close to) releasing large L3 datasets
    - Will be interesting to see the usage of these dataset
    - Releases broadly in line with expectation from the policy
    - In some cases the model will need to evolve to limit the huge dataset sizes
      - LHCb Ntuple Wizard, ALICE releasing derived/skimmed formats etc…
  - Experiments continuing to release Level-1 and Level-2 Open Data as before
- Expanding the set of experiments with formal Open Data policies
  - All small-LHC exeriments have now endorsed the policy
    - Will start to release L3 data before 5 years after the end of Run 3
  - Discussions ongoing with non-LHC experiments
    - Depending on the facilty:
      - Plan to write their own policy (tailored for their community)
      - Hope to sign up to the existing LHC policy
- Discussions ongoing about KPIs to monitor Open Data usage:
  - KPIs to be made public as part of the CERN Open Science reports
  - Internal KPIs for more detailed monitoring

Many thanks to the WG members for their valuable input during this process

Backup…

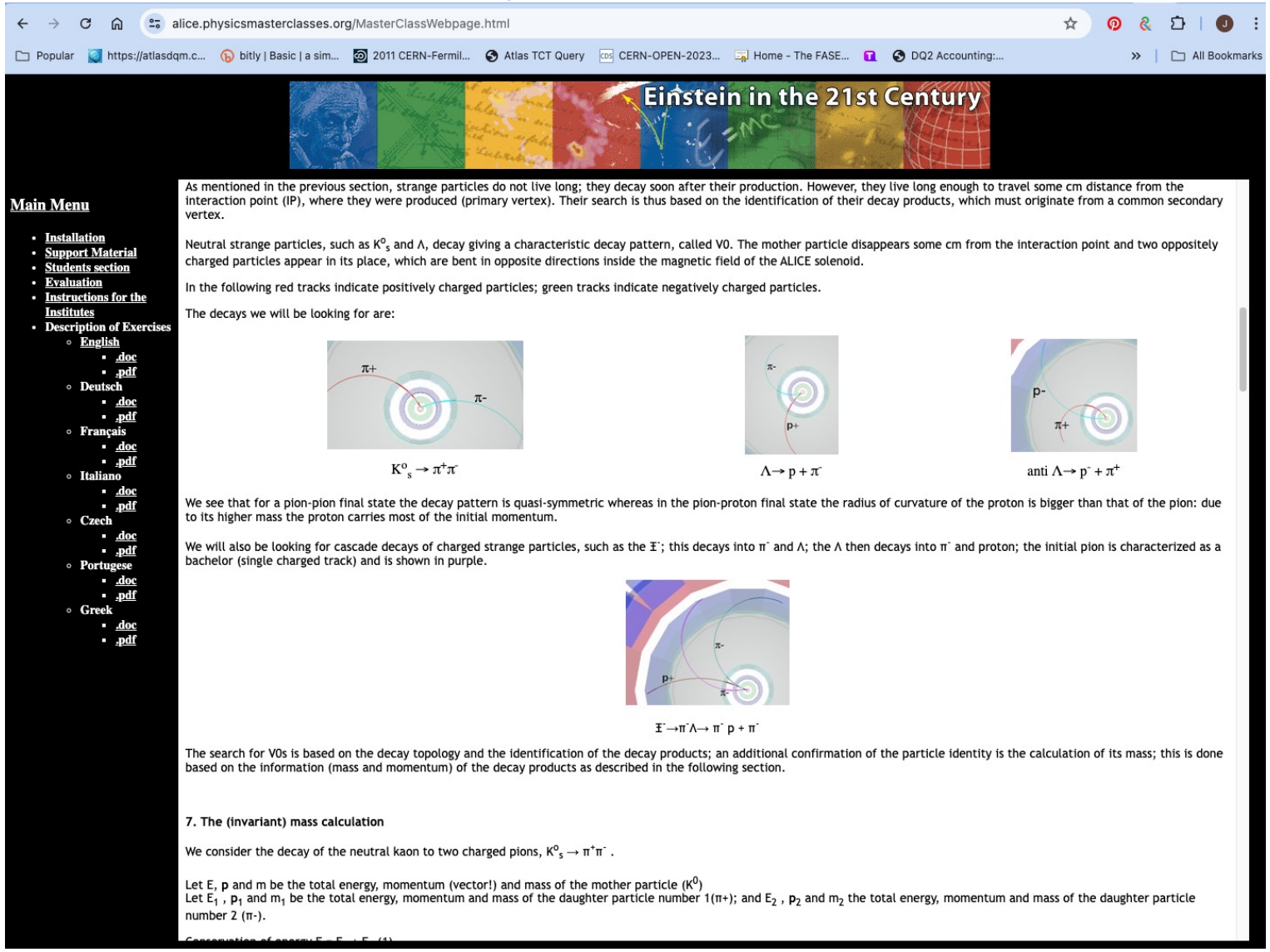
Level 1 Open Data

The HEPData is the tool for storing additional Level-1 data associated to a particular publication.
It can store digitized versions of plots, and more detailed information on event selections, efficiencies etc…



https://www.hepdata.net/

Level 2 Open Data



Example use of real experimental data for education and outreach. In this ALICE masterclass, school students can apply selections to real ALICE data, to emulate a published physics analysis.
The dataset used is openly available for education purposes.
The other large LHC experiments have similar tools for education purposes.

https://alice.physicsmasterclasses.org/MasterClassWebpage.html

Level 3 Open Data: LHCb

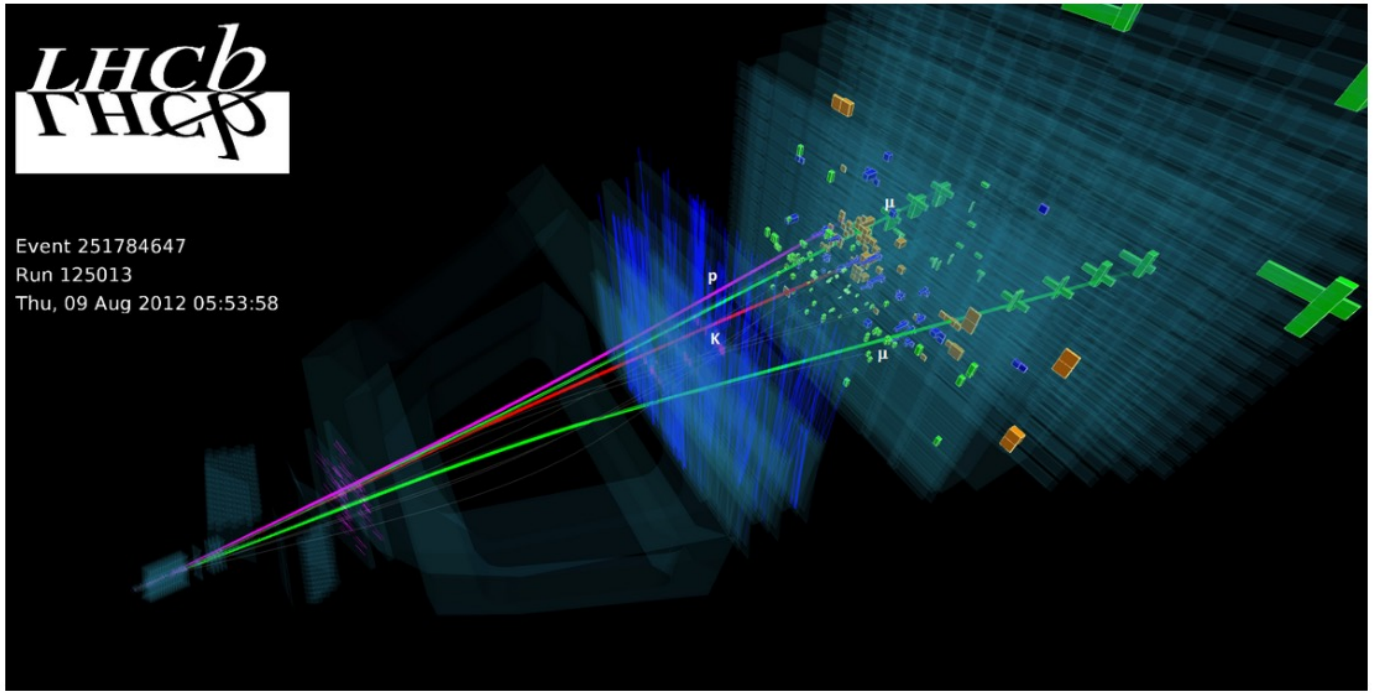
20/12/23

LHCb releases the entire Run I dataset

2023-12-20 by LHCb Collaboration

News

Today the LHCb collaboration completes the release of the data collected throughout the Run I of the Large Hadron Collider at CERN. The sample made available amounts to approximately 800 terabytes (TB) of data. These data, collected by the LHCb experiment in 2011 and 2012, contains information obtained from proton-proton collisions. The format made available provides pre-filtered data, suitable for a wide range of physics studies. The image below displays an event recorded during 2012.



Level 3 Open Data: CMS

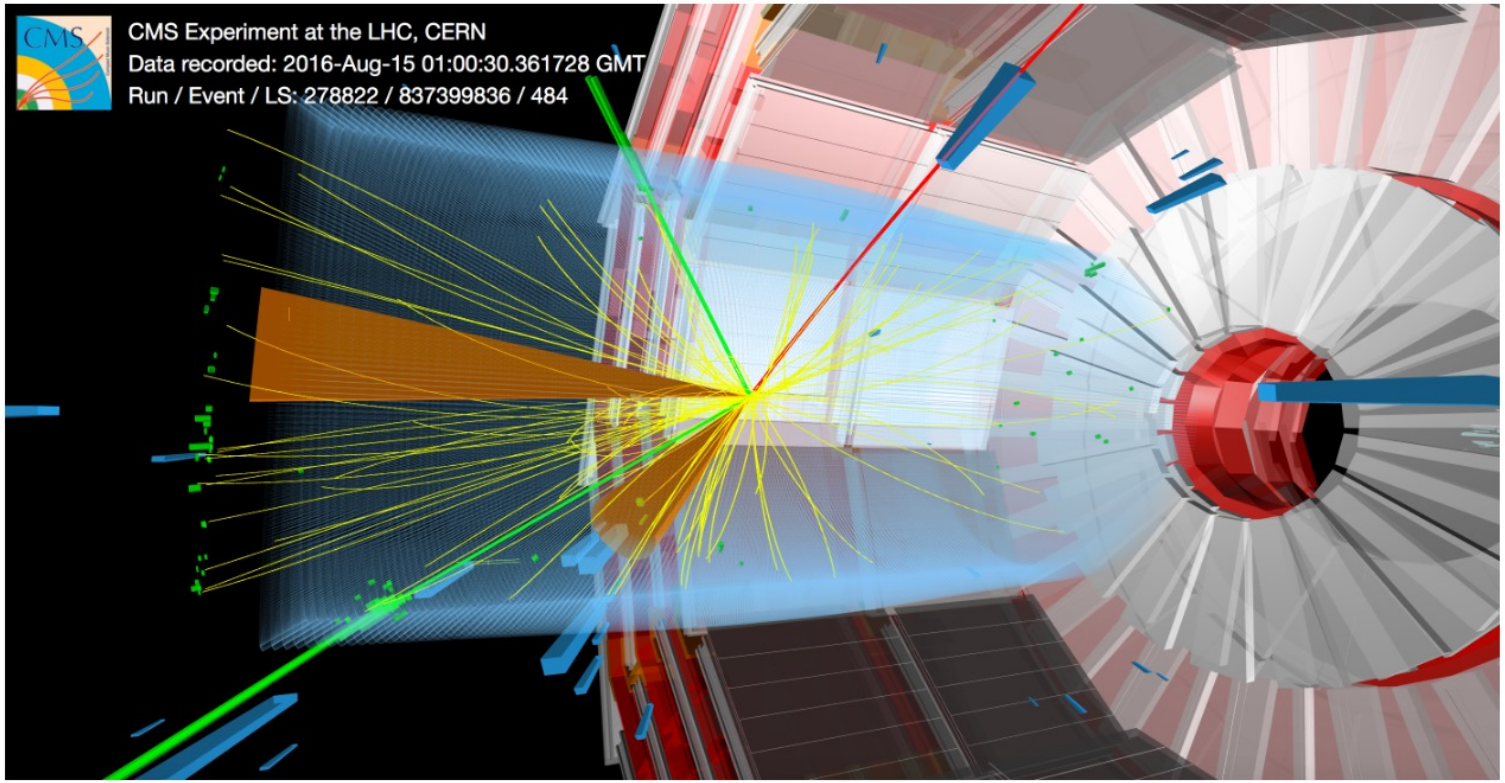
2/4/24

CMS releases 13 TeV proton collision data from 2016

2024-04-02 by CMS Collaboration

News

The CMS experiment at CERN is proud to announce the first release of 13 TeV proton-proton collision data collected in 2016. Over 70 TB of 13 TeV collision data and 830 TB of corresponding simulations are now accessible to the global scientific community and enthusiasts alike through the CERN Open Data Portal.



A candidate event in which a top quark is produced in association with a Z boson. Photo: CERN Document Server

Level 3 Open Data: ATLAS

1/7/24

ATLAS releases 65 TB of open data for research

2024-07-01 by ATLAS Collaboration

News

Explore over 7 billion LHC collision events – from home

The ATLAS Experiment at CERN has made two years' worth of scientific data available to the public for research purposes. The data include recordings of proton–proton collisions from the Large Hadron Collider (LHC) at a collision energy of 13 TeV. This is the first time that ATLAS has released data on this scale, and it marks a significant milestone in terms of public access and utilisation of LHC data.

"Open access is a core value of CERN and the ATLAS Collaboration," says Andreas Hoecker, ATLAS Spokesperson. "Since its beginning, ATLAS has strived to make its results fully accessible and reusable through open access archives such as arXiv and HepData. ATLAS has routinely released open data for educational purposes. Now, we're taking it one step further — inviting everyone to explore the data that led to our discoveries."

Released under the Creative Commons CC0 waiver, ATLAS has made public all the data collected by the experiment during the 2015 and 2016 proton–proton operation of the LHC. This is approximately 65 TB of data, representing over 7 billion LHC collision events. In addition, ATLAS has released 2 billion events of simulated "Monte Carlo" data, which are essential for carrying out a physics analysis.

External researchers, in particular, are encouraged to explore the ATLAS open data. "Along with the data, we have provided comprehensive documentation on several of our analyses, guiding users through our process step-by-step," says Zach Marshall, ATLAS Computing co-Coordinator. "These guides provide first-hand experience of working on a real ATLAS result, allowing anyone to test our tools, and evaluate the systematic uncertainties associated with the result for themselves."

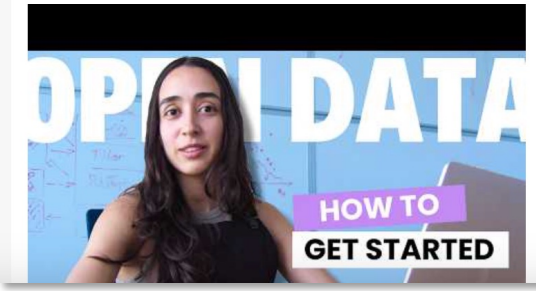
ATLAS traditionally collaborates with non-ATLAS scientists through short-term associations, granting them full access to ATLAS data, internal tools, and information. Through the open data, ATLAS researchers hope to further nurture this dialogue and collaboration. "In particular," adds Zach, "we'd like to encourage phenomenologists and also computer scientists to explore our datasets, instead of relying on mock-ups."

Today's release builds upon previous open data releases for educational use (in 2016 and 2020). "All of our open data releases are now available through the ATLAS open data website," says Dilia Portillo, ATLAS Outreach and Education co-Coordinator. "The website includes multi-level documentation, video tutorials and online tools aimed at the full-spectrum of users, from high school students to senior particle physics researchers. In addition, the software used to create the education-use open data has been released. This provides a seamless transition from the research open data to all the tutorials for outreach and education, including newly updated Higgs-boson discovery documentation. With a bit of time and dedication, you can go from being a relative novice to carrying out your own analysis."

The ATLAS open data website also serves as a hub for the community, which includes teachers, students, enthusiasts and, now, scientists. Anyone diving into the open data can also directly engage with ATLAS physicists, who are available to respond to user feedback and take suggestions.

This release marks the start of more to come, with ATLAS' first release of lead-lead-nuclei collision data data up next. The ATLAS Collaboration, along with the other main LHC experiment collaborations, has committed to making all of its data publicly accessible after a certain time. Openness is deeply ingrained in the culture of high-energy physics, enabling greater accessibility, reproducibility and better science.

Begin your journey with ATLAS open data by following the tutorial below.



WG Original Mandate

The European Commission is driving an Open Data Policy. CERN Council have encouraged the CERN experiments to make their data openly available for external analysis. The CERN Open Data portal is an initial attempt at making the data of the four experiments available through a common interface. So far this has largely been an experimental effort and a common strategy for Open Data has not been established.

This memo calls for forming a small **working group** to explore the conditions for a common Open Data policy from the LHC experiments. The **LHC collaborations are asked to nominate two representatives** each who will be complemented by one representative of RCS-SIS and IT each. The meetings will be convened by Jamie Boyd who reports to the Director of Research and Computing. The charge of the group is to draft a concise document on a common approach by the start of the summer addressing the following open points:

- Who are the expected users of the Open Data and what are their use cases?
    - What would be the target typical event size for the data that is made available and what would a rough conceptual breakdown of the event content look like? What operations (tightening selections, re-calibrating objects, applying systematic uncertainties etc.) would be possible / not-possible with this event format?
    - Would simulated data samples be made available with the data? If so which samples, and how would the total simulated data event size compare with the real data? Would there be a mechanism and/or recommendation for users to be able to obtain new simulated samples (e.g. of hypothetical signals)?
    - What would be the typical total annual data size on the Open Data portal? (note for this to be useful this should be something that can be downloaded to, and processed at a typical university computing installation);
- After what latency period would the collaborations relinquish their exclusive access to experimental data?
- Should there be any rules or recommendations related to who can use the Open data, for example related to members of the experiment that the data is from, or competing experiments?

L3 data – discussion - 2

- Summary of each experiments Latency shown below:

	ALICE	**ATLAS**	**CMS**	**LHCb**
Fraction of data released in: 5 yrs (6 yrs for CMS)	10%	25% (but limiting to <20% of the total data at that time)	50% (but limiting to <20% of the total data at that time)	50%
Fraction of data released in: 10 yrs	50%	50% (but limiting to <20% of the total data at that time)	100% (but limiting to <20% of the total data at that time)	100%
End-of-Collaboration	100%	100%	100%	100%

Resource needs (L3 data)

Expected disk resources for Open Data release (note LHCb have different trigger stratgey that increases the size of their dataset. This can be controlled by not releasing some of the exclusive streams if needed).

	ALICE	**ATLAS**	**CMS**	**LHCb**
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	**6 PB**	**1.5 PB**	**6 PB**	**55 PB**

It is expected that the computing resources will be covered by CERN IT.
The released data will be availabel through the CERN Open Data Portal.

The experiments will release data in their internal data-formats thus minimizing needed human resources.
The documents have a disclaimer that support for Open Data users will only be provided on a best-effort basis.

Slide from L. Heinrich

Text in public policy document on releasing L3 data

To be released:

- calibrated data
- accompanying Monte Carlo
- Analysis software
- Software environments (VM, containers)

Reconstructed Data (Level 3) Policy: The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies. The release of these data will be accompanied by provenance metadata, and by a concurrent release of appropriate simulated data samples, software, reproducible example analysis workflows, and documentation. Virtual computing environments that are compatible with the data and software will be made available. The information provided will be sufficient to allow high-quality analysis of the data including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A limited level of support for users of the Level 3 Open Data will be provided on a best-effort basis by the collaborations.

need for sufficient quality in OD

commit only to best-effort support

Public data releases will occur periodically, following an appropriate latency period to allow thorough understanding of the data, the reconstruction and calibrations, as well as to allow time for the scientific exploitation of the data by the collaboration. The size of the released datasets will be commensurate with the total amount of data collected of similar type, with the aim to commence data releases within five years of the conclusion of the run period. Data may be withheld by an experiment if there are active analyses ongoing. Full datasets will be made available at the close of the collaboration.

importance of sufficient latency period.

- start 5y after run period
- guideline:
  more data → more open data

- full data at close of expt
- escape hatch for ongoing analyses

The data will be released from the CERN Open Data Portal under the Creative Commons CC0 waiver, and will be identified with persistent data identifiers, and the data must be cited through these identifiers. Similarly, appropriate acknowledgements of the experiment(s) should be included in publications released using such data, and the publications made clearly distinguishable from those released by the collaboration. Any scientific claims in such publications are the responsibility of their authors and not of the experiments. It is expected that scientific results released using Open Data follow best scientific practices. The experiments may impose rules related to the use of the data by members of their respective collaborations.

data citation (of course) required

but waive responsibility for results

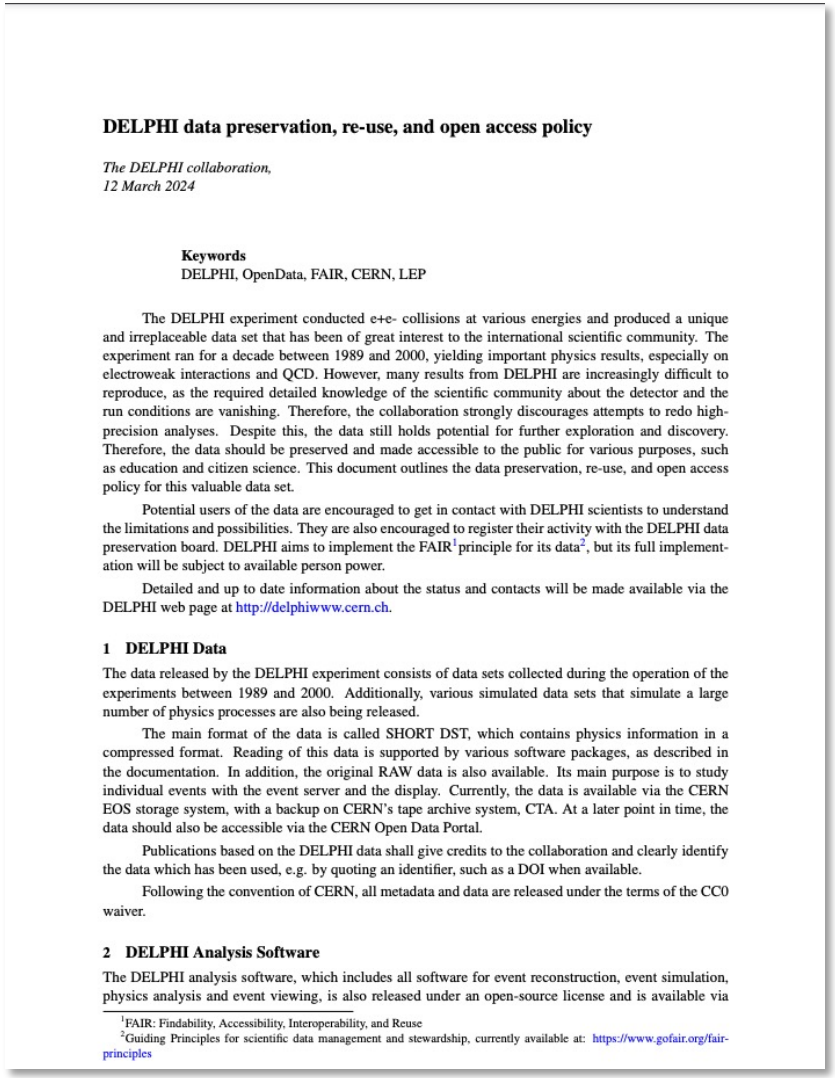
must be clearly marked as external publication

rules for use for own collaborators possible (but not required)

External authors should be aware that they will not have access to the vast amount of tacit knowledge built up within the LHC collaborations over the decades of design, construction and operation of the experimental apparatus. To allow external scientists to fully benefit from all the data, knowledge and tools, the collaborations may offer appropriate association programmes.

emphasize know-how asymmetry and advertise association programs

Expanding the effort – LEP expts



DELPHI data preservation, re-use, and open access policy

The DELPHI collaboration,
12 March 2024

Keywords
DELPHI, OpenData, FAIR, CERN, LEP

The DELPHI experiment conducted e+e- collisions at various energies and produced a unique and irreplaceable data set that has been of great interest to the international scientific community. The experiment ran for a decade between 1989 and 2000, yielding important physics results, especially on electroweak interactions and QCD. However, many results from DELPHI are increasingly difficult to reproduce, as the required detailed knowledge of the scientific community about the detector and the run conditions are vanishing. Therefore, the collaboration strongly discourages attempts to redo high-precision analyses. Despite this, the data still holds potential for further exploration and discovery. Therefore, the data should be preserved and made accessible to the public for various purposes, such as education and citizen science. This document outlines the data preservation, re-use, and open access policy for this valuable data set.

Potential users of the data are encouraged to get in contact with DELPHI scientists to understand the limitations and possibilities. They are also encouraged to register their activity with the DELPHI data preservation board. DELPHI aims to implement the FAIR[1] principle for its data[2], but its full implementation will be subject to available person power.

Detailed and up to date information about the status and contacts will be made available via the DELPHI web page at http://delphiwww.cern.ch.

1 DELPHI Data

The data released by the DELPHI experiment consists of data sets collected during the operation of the experiments between 1989 and 2000. Additionally, various simulated data sets that simulate a large number of physics processes are also being released.

The main format of the data is called SHORT DST, which contains physics information in a compressed format. Reading of this data is supported by various software packages, as described in the documentation. In addition, the original RAW data is also available. Its main purpose is to study individual events with the event server and the display. Currently, the data is available via the CERN EOS storage system, with a backup on CERN's tape archive system, CTA. At a later point in time, the data should also be accessible via the CERN Open Data Portal.

Publications based on the DELPHI data shall give credits to the collaboration and clearly identify the data which has been used, e.g. by quoting an identifier, such as a DOI when available.

Following the convention of CERN, all metadata and data are released under the terms of the CC0 waiver.

2 DELPHI Analysis Software

The DELPHI analysis software, which includes all software for event reconstruction, event simulation, physics analysis and event viewing, is also released under an open-source license and is available via

[1] FAIR: Findability, Accessibility, Interoperability, and Reuse
[2] Guiding Principles for scientific data management and stewardship, currently available at: https://www.gofair.org/fair-principles

LEP data represents the highest energy e+e- collision data, very valuable – especially in the context of future e+e- machines (FCC-ee etc…)

ALEPH:

- Open Data available and used in some physics papers
  - (not on CERN OpenData portal though)

DELPHI:

- Recently released Open Data policy
- Currently uploading data to CERN Open Data portal

- Status for L3/OPEL not clear, but probably not happening…

https://delphi-www.web.cern.ch/delphi-www/delsec/finalrules/DELPHI_Data_preservation-8.pdf
https://dphep.web.cern.ch/experiment/aleph

CMS Open Data Results

- CMS have been a pioneer of releasing L3 Open Data
- There have been a number of publications on this data covering BSM, QCD, jet-reconstruction and machine learning topics