LHC Open Data

on behalf of the experiments at CERN

Dr. Mindaugas Šarpis

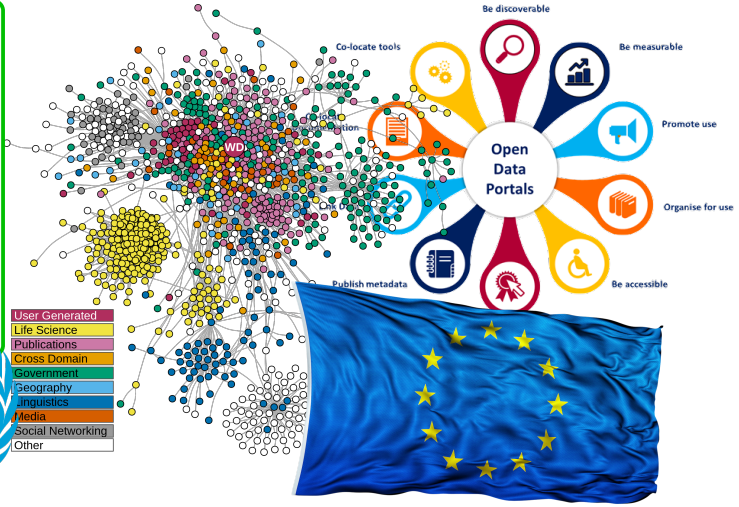
Vilnius University / The University of Manchester

June 3, 2024

The Idea of Open Data

Key principles of Open Data:

- Freely available to use, reuse and redistribute
- Enhances transparency
- Enhances social impact
- Enables involvement of 3rd parties
- Enables future research innovation



F.A.I.R. Principles - Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- **F1.** (Meta)data are assigned a globally unique and persistent identifier
- **F2.** Data are described with rich metadata (defined by R1 below)
- **F3.** Metadata clearly and explicitly include the identifier of the data they describe
- **F4.** (Meta)data are registered or indexed in a searchable resource

[Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)]

F.A.I.R. Principles - Accessible

Once the user finds the required data, it must be clear how they can be accessed, possibly including authentication and authorization.

- **A1.** (Meta)data are retrievable by their identifier using a standardized communications protocol
  - **A1.1** The protocol is open, free, and universally implementable
  - **A1.2** The protocol allows for an authentication and authorization procedure, where necessary
- **A2.** Metadata are accessible, even when the data are no longer available

[Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)]

F.A.I.R. Principles - Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- **I1.** (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **I2.** (Meta)data use vocabularies that follow FAIR principles
- **I3.** (Meta)data include qualified references to other (meta)data

[Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)]

F.A.I.R. Principles - Reusable

The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- **R1.** (Meta)data are richly described with a plurality of accurate and relevant attributes
  - **R1.1.** (Meta)data are released with a clear and accessible data usage license
  - **R1.2.** (Meta)data are associated with detailed provenance
  - **R1.3.** (Meta)data meet domain-relevant community standards

[Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016)]

CERN Level 3 Open Data Release Policy

LHCb and other experiments at CERN have made a decision to make the collected data available to the public.

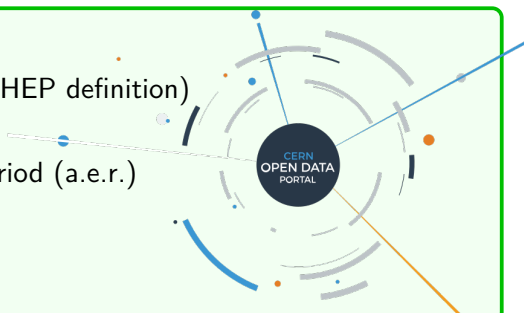
This should be done in-line with **F.A.I.R.** principles.

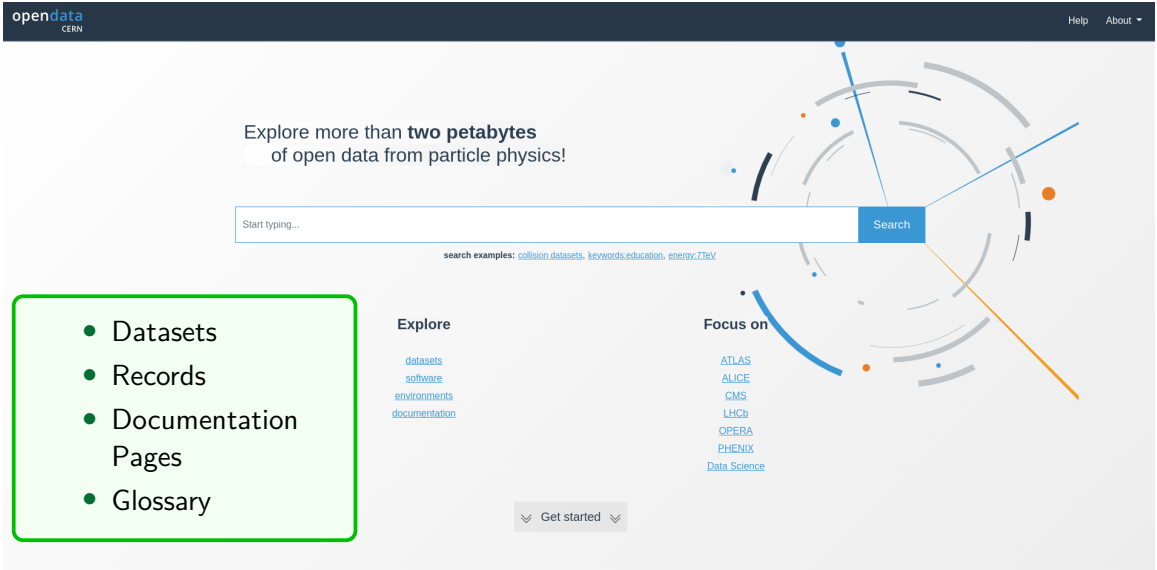
Which data to release?

- Level 3 open data: reconstructed events (DPHEP definition)

When to release it?

- 50 % of data 5 years after end of running period (a.e.r.)
  - Run I: End of 2017
- 100 % of data 10 years a.e.r.
  - Run I: End of 2022





- Datasets
- Records
- Documentation Pages
- Glossary

LHCb 2012 Beam4000GeV MagDown EW Stream Stripping21

LHCb collaboration

Dataset | Collision | LHCb | EW | CERN-LHC

Description

proton-proton (pp) collision data collected by the LHCb experiment in the year 2012 of Run1 of the LHC.

Dataset characteristics

334776609 events, 6292 files, 19.3 TB in total.

How were these data selected?

This dataset was created in several production steps. These steps, software used and the configuration is provided below.

Prod ID: 41836

Prod type: Merge

Parent Prod ID: 41835

Parent Prod type: DataStripping

Conditions:

dddb-20130929-1

cond-20141107

List of Trigger Configuration Keys (TCK):

TCK	Number of Files
0x860040	42
0x7f0040	4
0x8c0040	699
0x7e003a	9
0x94003d	570
0x95003d	11
0x97003d	774
0x8e0040	3
0x990042	1091
0x990044	633
0xa30044	935
0xa90046	416
0xa30046	60
0xab0046	379
0xac0046	634
0xad0046	32

Dataset × Collision × LHCb × 2011 × DST × pp × MagDown × EW × stripping21r1 ×

include on-demand datasets

Filter by type
- Dataset 1
- Collision 1

Filter by experiment
- LHCb 1

Filter by year
- 2011 1

Filter by file type
- DST 1

Filter by collision type
- pp 1

Filter by collision energy
- 7TeV 1

Filter by magnet polarity
- MagDown 1
- MagUp 1

Filter by stripping stream
- EW 1
- RADIATIVE 1

Filter by stripping version
- stripping21r1 1
- stripping21r1p1 1
- stripping21r1p2 1

Sort by: Best match, asc. Display: detailed, 20 results

Found 1 result.

LHCb 2011 Beam3500GeV MagDown EW Stream Stripping21r1

proton-proton (pp) collision data collected by the LHCb experiment in the year 2011 of Run1 of the LHC...

Dataset | Collision | LHCb

« 1 »

Code	(MINTREE('mu+'==ABSID,PT) > 650.0 *MeV) & (MINTREE('mu+'==ABSID,P) > 8000.0 *MeV) & (MAXTREE('mu+'==ABSID,TRCHI2DOF) < 5.0) & (MM > 3000.0) & (VFASP
Inputs	['Phys/ StdLooseDiMuonSameSign ']
DecayDescriptor	None
Output	Phys/MicroDSTDiMuonDiMuonSameSignLine/Particles

ABSID

evaluator of the absolute value for the particle id
const LHCb::ParticleProperty * ...
int pid = (int) ABSID(p) ;
See also
LoKi::Particles::AbsIdentifier
LHCb::ParticleID
LHCb::IParticle

Author
Vanya Belyaev Ivan.Belyaev@itep.ru

Date
2002-07-15

Definition at line 133 of file ParticleCuts.h.

LHCb Stripping V21 commonparticles

Documentation Stripping

stripping21

Common particles:

- Basic
- Intermediate

Standard basic particles:

- StdAllLooseANNElectrons
- StdAllLooseANNKaons
- StdAllLooseANNPions
- StdAllLooseANNProtons
- StdAllLooseElectrons
- StdAllLooseGammaDD
- StdAllLooseGammaLL
- StdAllLooseKaons
- StdAllLooseMuons
- StdAllLoosePions
- StdAllLooseProtons
- StdAllNoPIDsElectrons
- StdAllNoPIDsKaons
- StdAllNoPIDsMuons
- StdAllNoPIDsPions
- StdAllNoPIDsProtons
- StdAllTightGammaDD
- StdAllTightGammaLL
- StdAllTightSymGammaDD
- StdAllTightSymGammaLL
- StdAllVeryLooseMuons
- StdDIElectronFromTracks
- StdDIElectronGamma
- StdJets
- StdLooseANNDownElectrons
- StdLooseANNDownKaons
- StdLooseANNDownPions
- StdLooseANNDownProtons
- StdLooseANNElectrons
- StdLooseANNKaons

LHCb Stripping V21 stdallloosegammall

Documentation Stripping

[stripping21 lines]

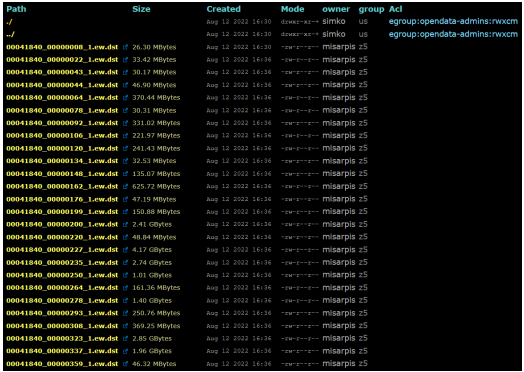
StdAllLooseGammaLL

DiElectronMaker/StdAllLooseGammaLL

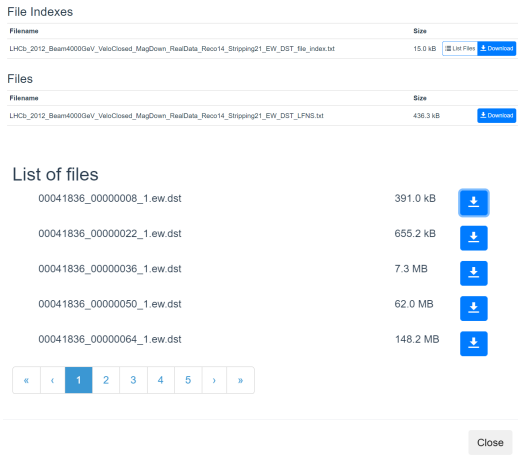
Inputs	[]
Input	Rec/ProtoP/Charged
DecayDescriptor	gamma -> e+ e-
Output	None
Particle	gamma
Tools:	

ProtoParticleCALOFilter/StdAllLooseGammaLL.Electron

AuditFinalize :	False
AuditInitialize :	False
AuditStart :	False
AuditStop :	False
AuditTools :	False
Context :	None
ContextService :	AlgContextSvc
CounterList :	['.*']
EfficiencyRowFormat :	' | %-48.48s | %50i | %10d | %11.5g | (%)#9.6g | +- %|-#9.6g|)| %| ------- | ------- | '
ErrorsPrint :	True
GlobalTimeOffset :	0.0



Path	Size	Created	Mode	owner	group	Acl
./		Aug 12 2022 16:30	drwxr-xr-x	simko	us	egroup:opendata-admins:rwxcm
~/		Aug 12 2022 16:30	drwxr-xr-x	simko	us	egroup:opendata-admins:rwxcm
00041840_00000008_1.ew.dst	26.30 MBytes	Aug 12 2022 16:30	-rw-r--r--	misarpis	z5	
00041840_00000022_1.ew.dst	33.42 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000043_1.ew.dst	30.17 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000044_1.ew.dst	46.90 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000064_1.ew.dst	370.44 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000078_1.ew.dst	30.31 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000092_1.ew.dst	331.02 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000106_1.ew.dst	221.97 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000120_1.ew.dst	291.43 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000134_1.ew.dst	32.53 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000148_1.ew.dst	135.07 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000162_1.ew.dst	625.72 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000176_1.ew.dst	47.19 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000199_1.ew.dst	150.88 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000200_1.ew.dst	2.41 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000220_1.ew.dst	48.84 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000227_1.ew.dst	4.17 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000235_1.ew.dst	2.74 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000250_1.ew.dst	1.01 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000264_1.ew.dst	161.36 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000278_1.ew.dst	1.40 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000293_1.ew.dst	250.76 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000308_1.ew.dst	369.25 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000323_1.ew.dst	2.85 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000337_1.ew.dst	1.96 GBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	
00041840_00000350_1.ew.dst	46.32 MBytes	Aug 12 2022 16:36	-rw-r--r--	misarpis	z5	



File Indexes

Filename	Size	
LHCb_2012_Beam4000GeV_VeloClosed_MagDown_RealData_Reco14_Stripping21_EW_DST_file_index.txt	15.0 kB	List Files Download

Files

Filename	Size	
LHCb_2012_Beam4000GeV_VeloClosed_MagDown_RealData_Reco14_Stripping21_EW_DST_LFNS.txt	436.3 kB	Download

List of files

00041836_00000008_1.ew.dst	391.0 kB
00041836_00000022_1.ew.dst	655.2 kB
00041836_00000036_1.ew.dst	7.3 MB
00041836_00000050_1.ew.dst	62.0 MB
00041836_00000064_1.ew.dst	148.2 MB

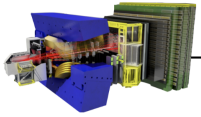
« ‹ 1 2 3 4 5 › »

Close

LHCb Run I Data Release

- Over **800TB** of files moved to dedicated storage
- Index files created for every data set
- List of LFNS stored for prevalence
- $\sim$ **7500** LHCb Stripping Pages converted to Open Data Portal format and provided with the rest of documentation
- Glossary of Open Data Portal is enriched with **960** LHCb specific terms (like LoKi functors)
- Routine for scraping metadata from `DIRAC` created
- LHCb Open Data already adopted for educational purposes

What about Run II (and beyond)?



5TB/s

Centralized Data Handling

Trigger
Moore

100GB/s

Reconstruction
Brunel

10GB/s

Stripping
DaVinci

15PB/year

Re-reconstruction

Restripping

Storage
FULLSTREAM

Storage
FULLSTREAM

Storage
CHARM.MDST
DIMUON.DST
...

Particle simulation
Gauss
MC

Digitisation
Boole
MC

Generation
Pythia/POWHEG/...

Decay
EvtGen

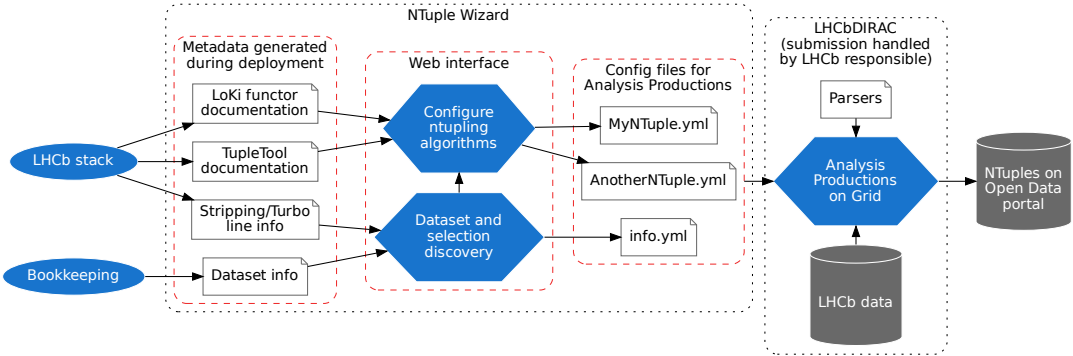
Propagation
Geant4

Analysis
ROOT/Numpy/...

Ntuple making
DaVinci

Personal Data Handling

What about Run II (and beyond)?

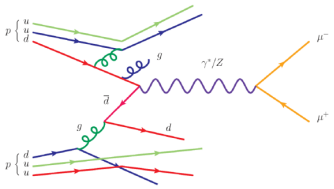


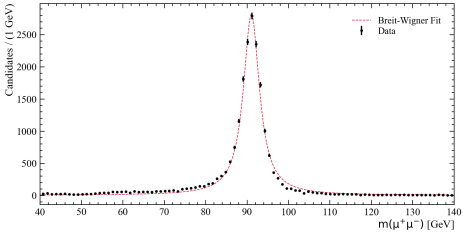
[Ntuple Wizard: An Application to Access Large-Scale Open Data from LHCb]

Using Open Data

[Using CMS Open Data in research – challenges and directions]

- Jet Substructure Studies:
  - [Jet Substructure Studies with CMS Open Data]
  - [Exposing the QCD Splitting Function with CMS Open Data]
- Searches for New Particles:
  - [Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum]
  - [Search for Non-Standard Sources of Parity Violation in Jets at $\sqrt{s} = 8\,TeV$ with CMS Open Data]
- Standard Model Analyses:
  - [Exploring Uncharted Soft Displaced Vertices in Open Data]





Open Questions

- How do we keep data open with increasing size of the datasets?
- How do we tackle the increasing complexity of particle physics data analysis wrt. open data?
- Can we start releasing raw data?
- How do we increase the focus on open data?
- How do we fund open data efforts?
- How usefull is LHC open data in the end?

End of slides

Thanks for your attention