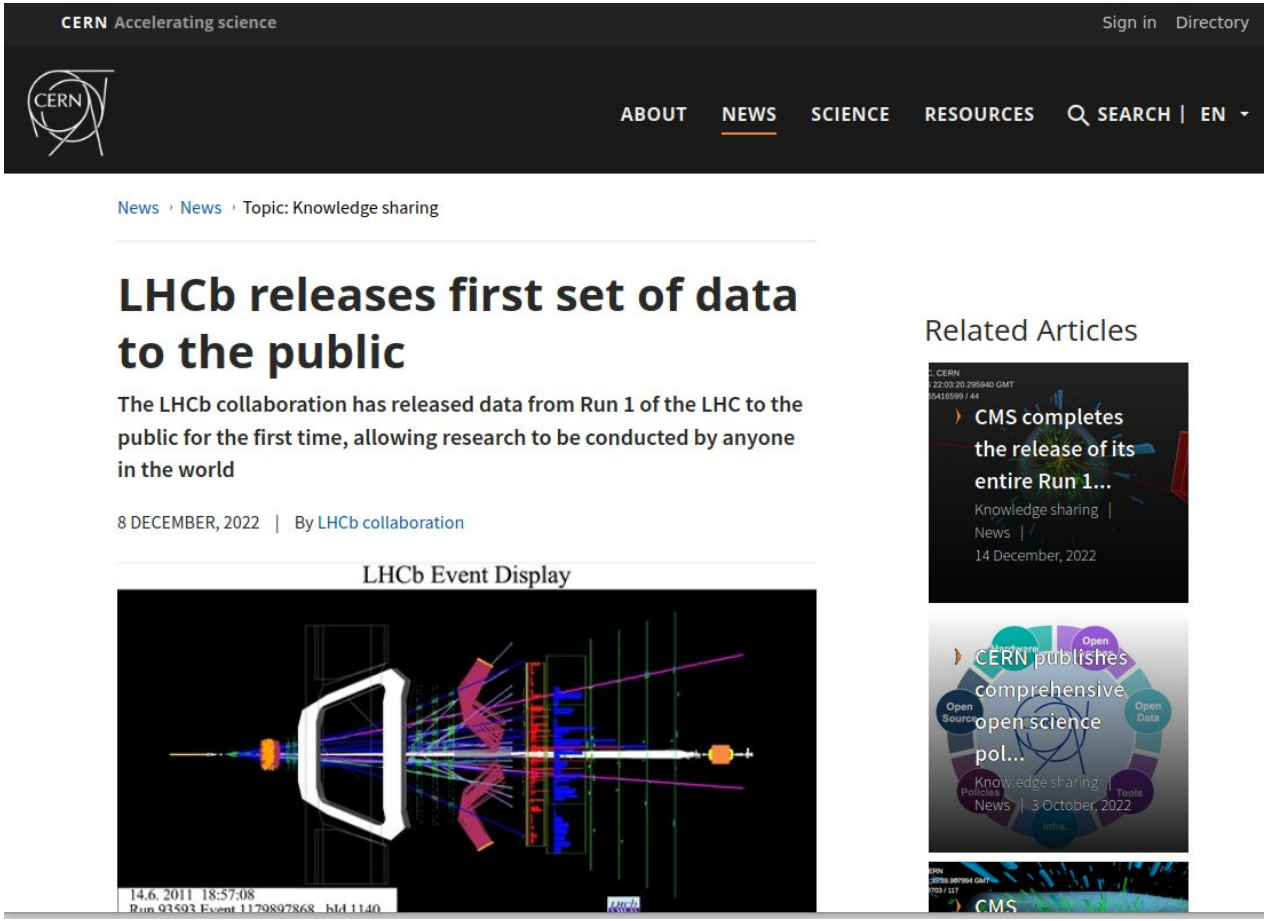
Open Data Activities at LHCb

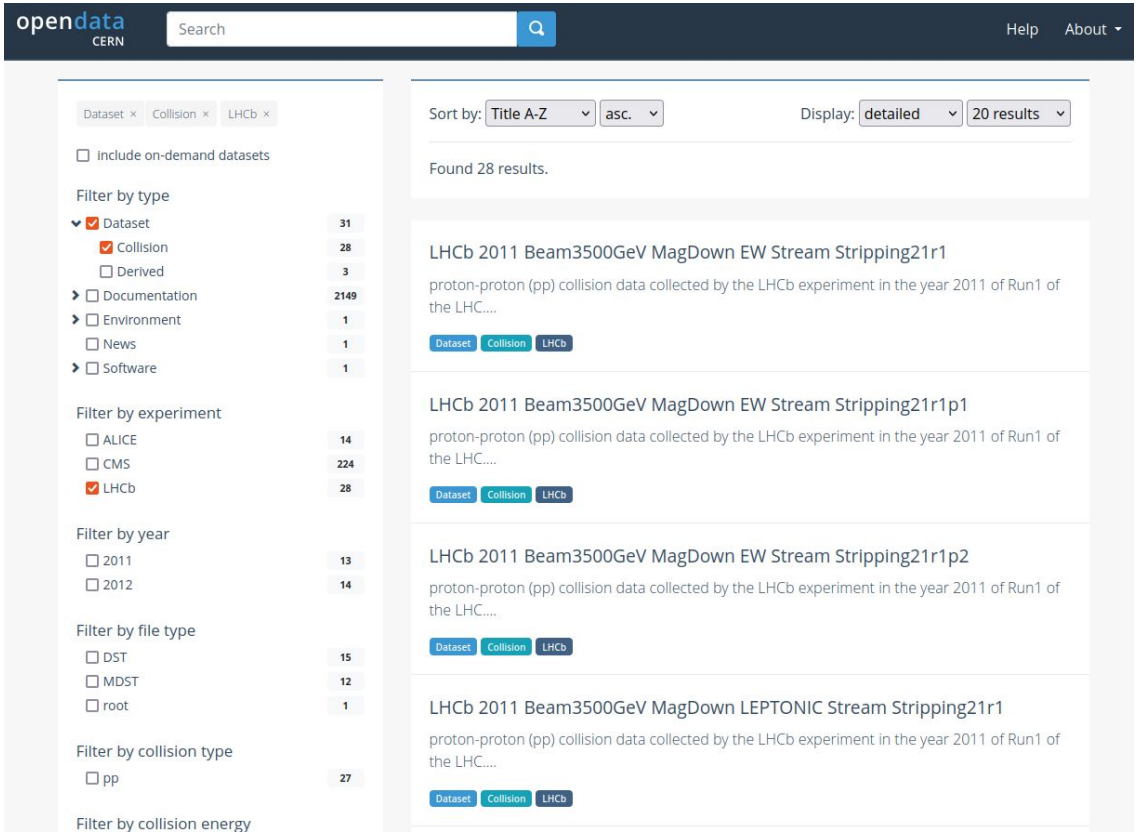
Dillon Fitzgerald[1] on behalf of the LHCb Collaboration
[1]University of Michigan

HSF Data Analysis WG,
Open Data Meeting
03.07.2023



https://home.cern/news/news/knowledge-sharing/lhcb-releases-first-set-data-public

LHCb Open Data Currently Available on the OD Portal



https://opendata.cern.ch/search?page=1&size=20&experiment=LHCb&subtype=Collision&type=Dataset

Goals of Open Data - Maximizing Scientific Value

- **Validation / reproduction of published results**
- **Reinterpretation of the data**
  - Test future theories
  - Refine phenomenological models
  - Use different statistical tools
- **Reuse of data sets**
  - Combined analyses
  - Use collected data as input for future studies
  - Algorithm development (e.g. machine learning community)
- **Data mining**
  - Search for interesting physics in unexplored parts of the data
  - Use new techniques to (re-)select data

We cannot anticipate the questions future generations might ask of this data.

require different levels of data complexity

Policies

the CERN experiments have given themselves

CERN Open Data Policy 2020

Initiated beginning 2020 by the chair of the European Commission

CERN director of research: Mandate for a working group to draft a common policy for all LHC experiments

Endorsed by the Collaboration Boards of ALICE, ATLAS, CMS and LHCb

CERN Open Science Policy 2022

Includes all experiments at CERN

Working group formed
https://openscience.cern/

Includes a wider scope of topics:

- Open access, open data, open source, open hardware
- Research integrity, research assessment
- Open infrastructure
- Training and outreach, citizen science

FAIR Data Principles

[The FAIR Guiding Principles for scientific data management and stewardship. Nature *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18]

Findable: Metadata and data should be easy to find for both humans and computers.

Accessible: The exact conditions under which the data is accessible should be provided in such a way that humans and machines can understand them.

Interoperable: The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.

Reusable: Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

https://go-fair.org

Solved by
https://opendata.cern.ch

Needs dedicated work by the experimental collaborations (here efforts in HEP are in their infancy)

Data Preservation in HEP (DPHEP) Levels of Data Complexity

[https://arxiv.org/abs/1205.4667]

1. **Published results + additional information**
   - Supplemental data tables, ntuples
   - HEPData entries, rivet plugins
   - Notes, technical information
   - Documentation, slides
   - Analysis code, jupyter notebooks

2. **Education and Outreach**
   - Simplified data formats, e.g. highly preprocessed ntuples

3. **Reconstructed data + analysis level software**
   - Calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies
   - Preservation of analysis level experiment-specific software

4. **Raw data + reconstruction software**
   - **Not released for LHC data**

Open data policy: Level 3 data releases

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.

Level 3 data is addressed at professional researchers

Level 3 open data release policy - LHCb

Policy since 27th Feb 2013 updated in CERN Open Data Policy 2020 and CERN Open Science Policy 2022

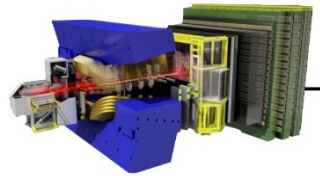
- 50% of data 5 years after end of running period (a.e.r.)
  - Run I: End of 2017
  - **Run II: End of 2023**
- 100% of data 10 years a.e.r.
  - Run I: End of 2022
- Level 3 open data: reconstructed events (DPHEP definition)
  - LHCb: Output of Stripping / Turbo / Sprucing
  - MC on demand
- Goal of the OD release is to enable scientific research by 3rd parties
- **Level 3 data releases are addressed at professional users**

This model differs from the astrophysics model of open data.

Reason: incentive structures for instrument/detector builders vs data analysts, differ widely between the two fields

LHCb Level 3 Data

Release policy: 50% @ 5yrs, 100% @ 10yrs after end of running period



- Level 3 data in LHCb defined as the output of the Stripping
- Same level of abstraction accessed by LHCb members
- Contains comprehensive set of selections (1620 selections in latest version)
- Organized in ~10 streams, according to physics signature
- Software needed to access data (DaVinci) is open source, available via CVMFS (or container)
- Documentation: LHCb Starterkit openly available

LHCb Run 1 open data release

- Released 3 Streams:
  - Electroweak EW,
  - Leptonic,
  - Radiative
- ~ 200TB (roughly 20% of Run 1 data)
- Data released in LHCb MDST and DST formats
- Needs DaVinci application to read

- Detailed description of Stripping selections
- Glossary of 960 LHCb specific terms
  - Mostly defining possible selection criteria

- Monte Carlo samples on demand

See slides 2 and 3 for the corresponding press release and data files on the CERN Open Data Portal

CERN Open Data releases are managed through open GitHub repository (for metadata):

https://github.com/cernopendata

See for yourself! https://opendata.cern.ch/

Resources needed to support open data

LHCb understaffed!

OD Release:

1 Senior + 1 PhD + CERN OD team

OD developments (Ntuple Wizard):

2 Seniors, 1 Postdoc, 2 PhDs, summer student + support from CERN and LHCb IT

For comparison CMS:

O(5) seniors + groups

\+ CMS IT + CERN OD team

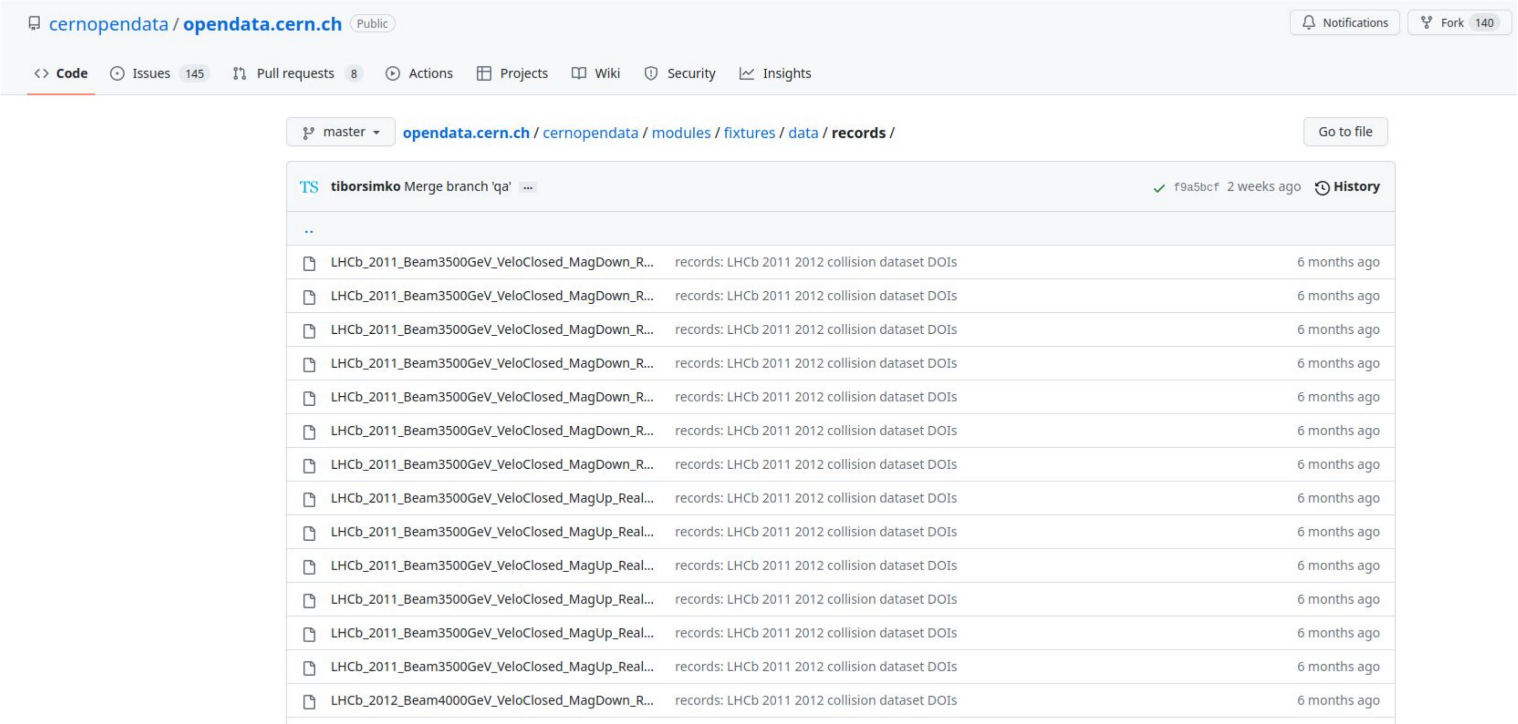
Curating open data requires significant effort

- Decision on datasets to be released
- Preparation of files
- Preparation of meta-data (!)
- Integration into OD portal

CERN Open data curation on GitHub

Many records have been prepared containing metadata on available datasets

- See an [example json record]



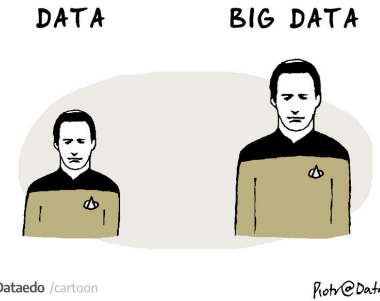
Level 3 Data - Resources

2020 Projections

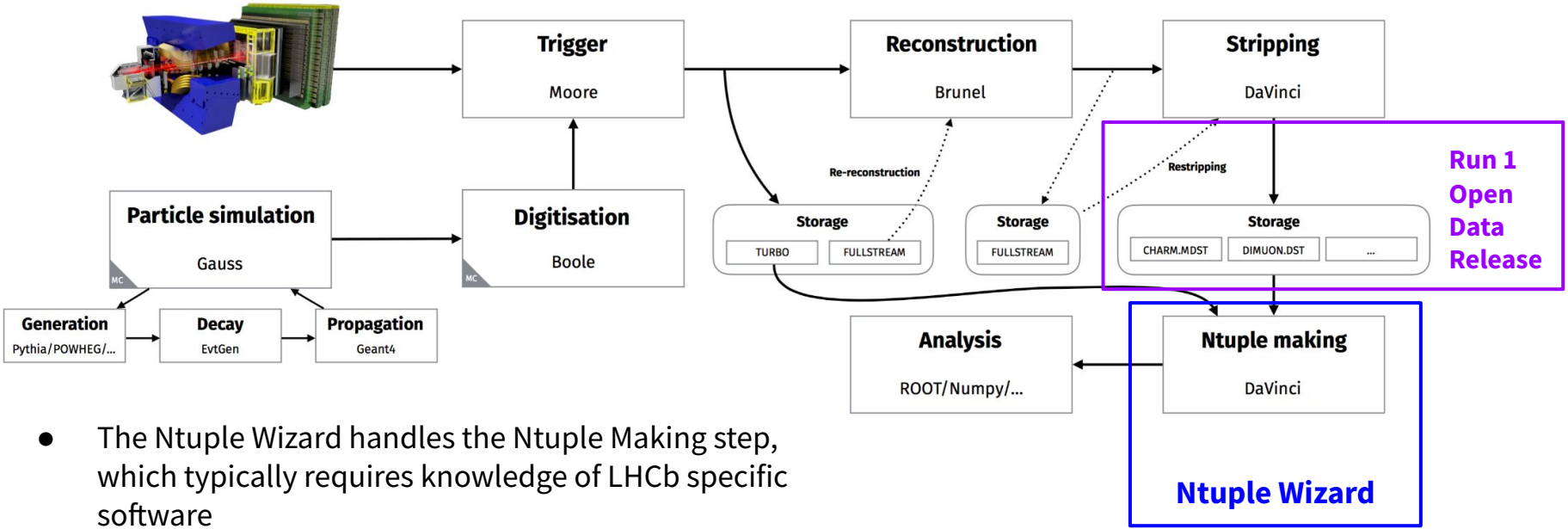
	ALICE	ATLAS	CMS	LHCb
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	**6 PB**	**1.5 PB**	**6 PB**	**55 PB** Not scalable!

Mitigation Strategies:

- Provide protected access to existing copies of Stripping/Turbo output via WG-production slots. **Needs a dedicated tool, enter: "Ntuple Wizard"**
- Provide direct access to data on grid storage

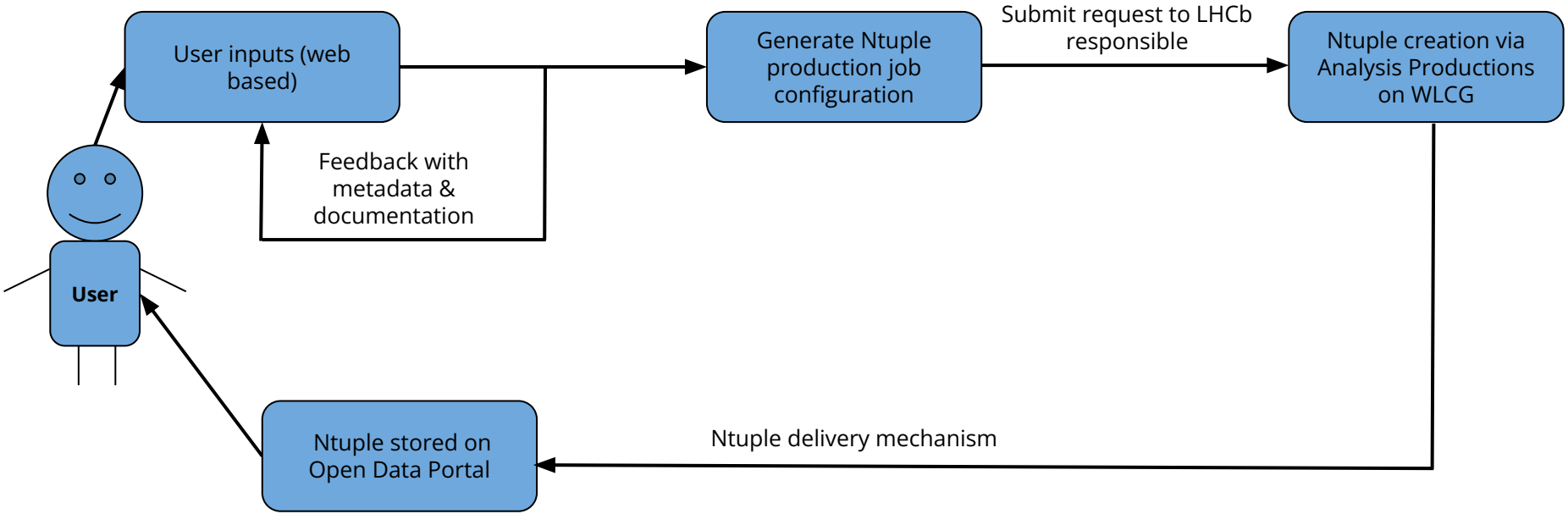


LHCb Level 3 Data: Ntuple Wizard

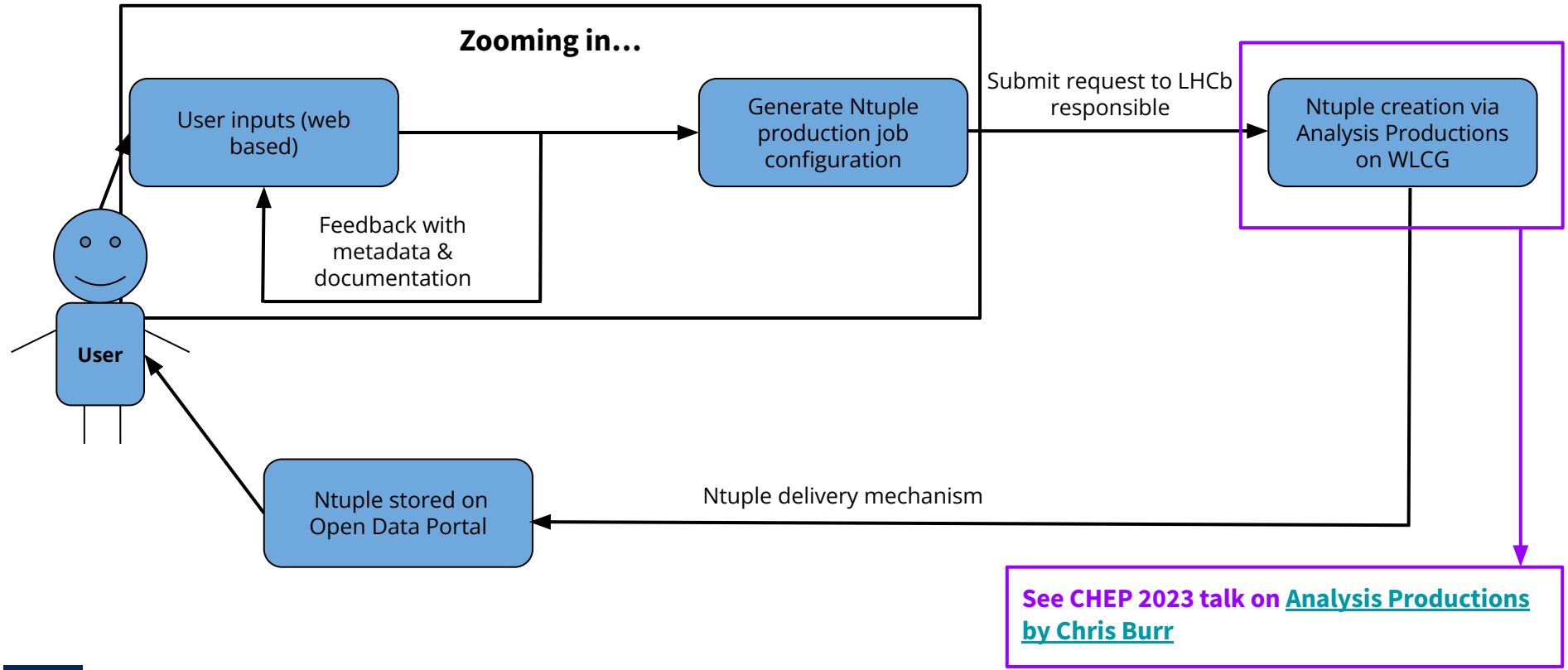


- The Ntuple Wizard handles the Ntuple Making step, which typically requires knowledge of LHCb specific software
  - **Lower barrier of entry for external analysts!**

The Ntuple Wizard



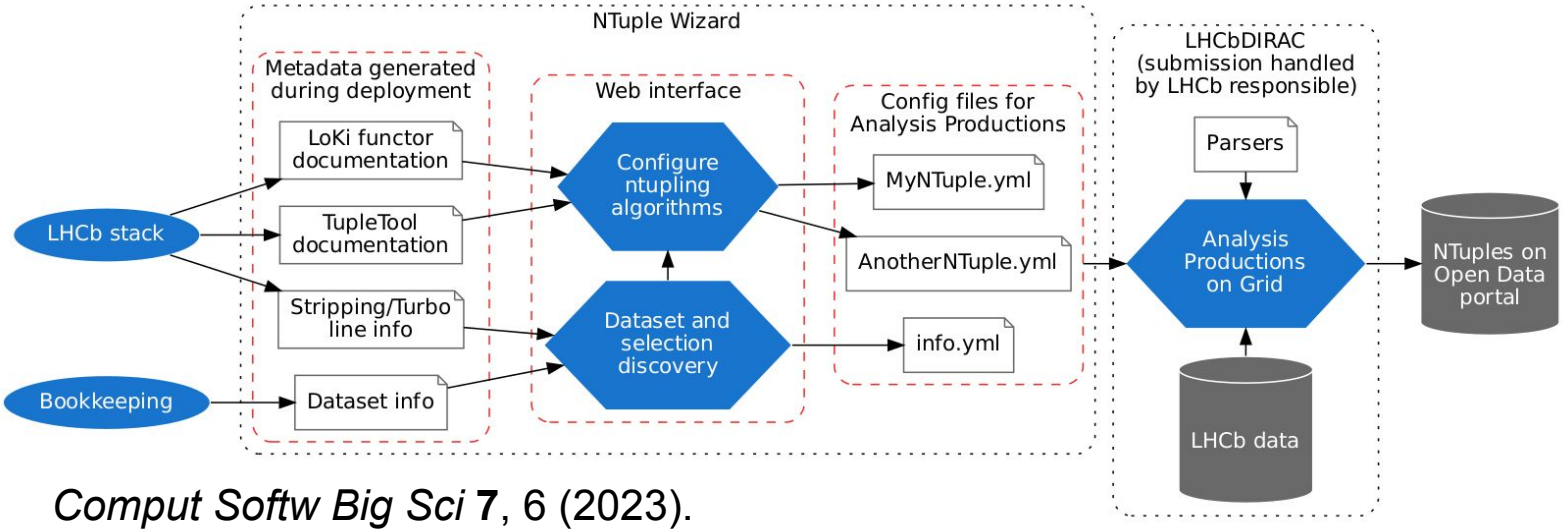
The Ntuple Wizard



The Ntuple Wizard

LHCb members can test and provide feedback on (mattermost)

- The Ntuple Wizard is functional https://lbwizard.web.cern.ch/
- **Idea:** only keep existing replicas of the data, allow OD users access via dedicated analysis production jobs
  - Metadata generation during deployment (input) and config files for Analysis Productions (output) facilitate secure design

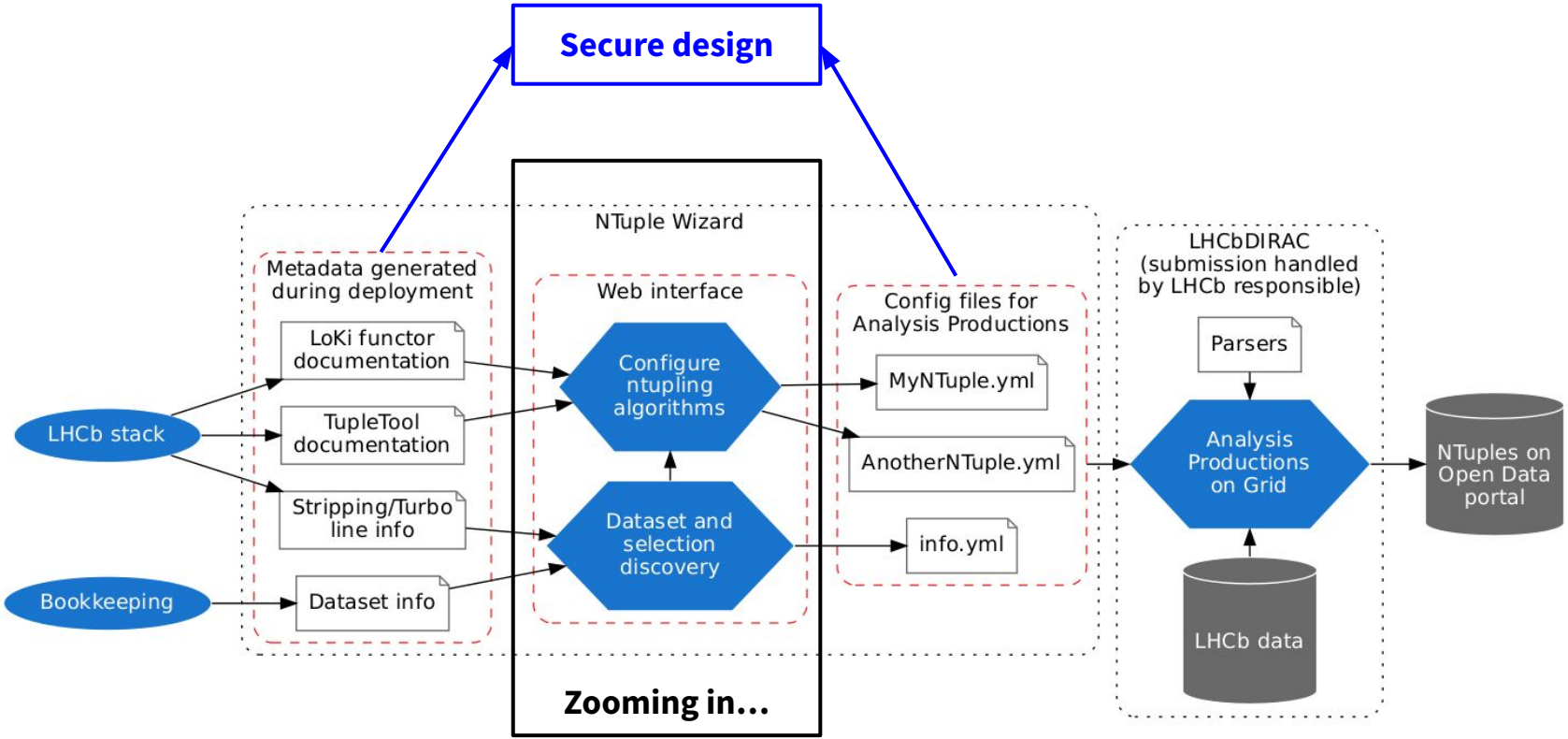


Comput Softw Big Sci **7**, 6 (2023).
https://doi.org/10.1007/s41781-023-00099-5

Interlude: Security & Permissions

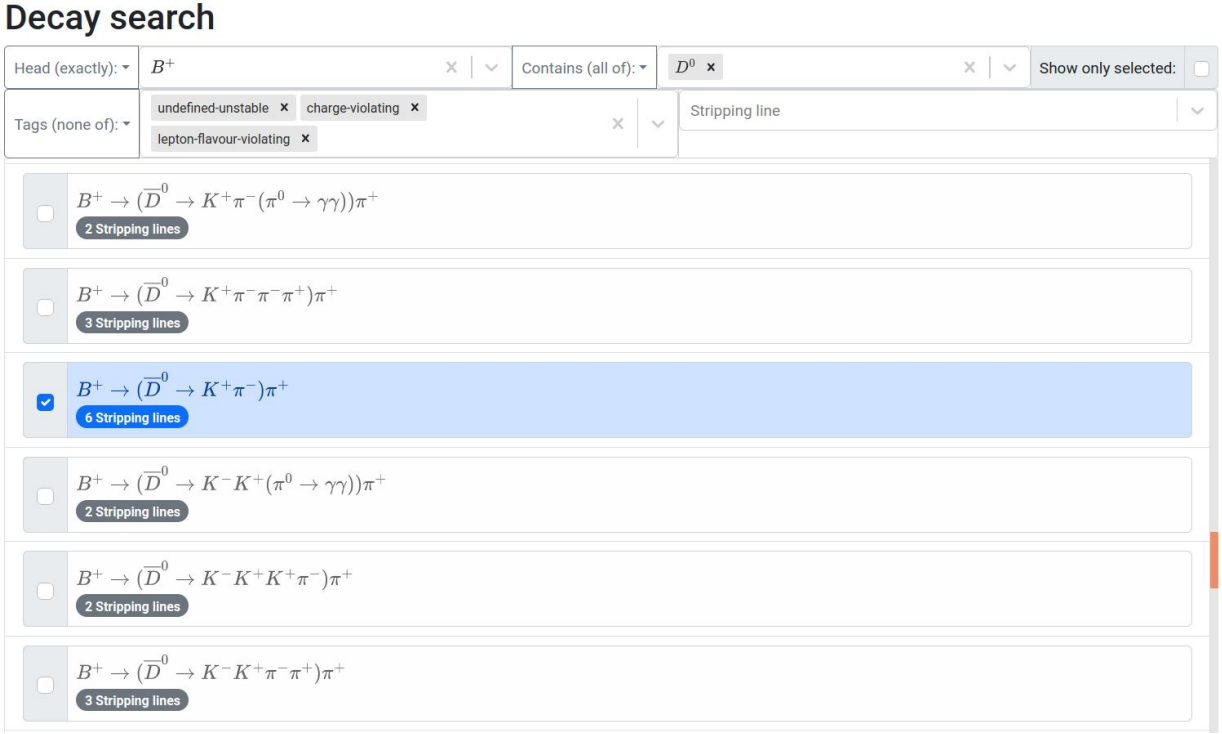
- Standard LHCb Ntuple making application (DaVinci) configured with python scripts
  - Running arbitrary code from external users is a security risk!
  - Config output saved in yaml data structures, interpreted by internal parsers
- Dataset discovery and Ntuple configuration require metadata from the LHCb database and software stack
  - Metadata is extracted at “deployment time”
  - Only static files are read at run time, no interaction with LHCb database system
- LHCb policy reserves right to withhold part of a dataset (e.g. ongoing analyses)
  - Run 1 open data release only contains a subset of the data because of this
  - The Ntuple Wizard can improve this via fine-grained control over:
    - Building/accessing decay candidates
    - Stripping lines or equivalent selections

The Ntuple Wizard



Web Interface: Dataset Discovery

***Key feature:** Find available dataset by first choosing physics object of interest!



Lists physics objects available in the LHCb database (primarily decays)

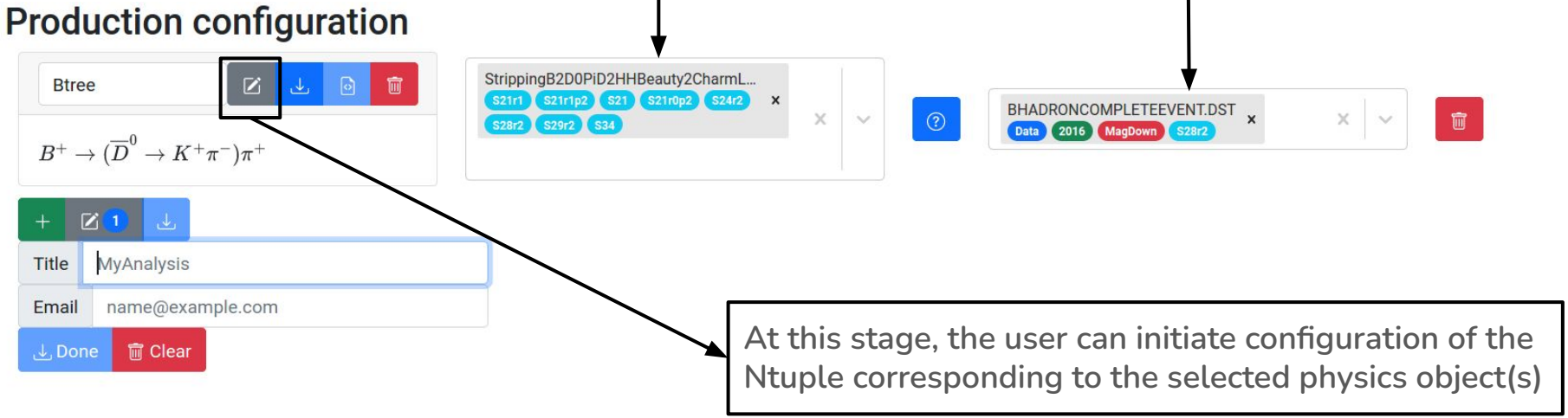
- List filtering options include:
  - Decay head (top level decaying particle)
  - Particles in the decay
  - Tags related to specific physics (include or exclude)
  - “Stripping line” name
    - more useful for LHCb internal users
- Can make multiple selections from the list

Web Interface: Dataset Discovery

Selection of a physics object exposes the corresponding available datasets for the user to choose from

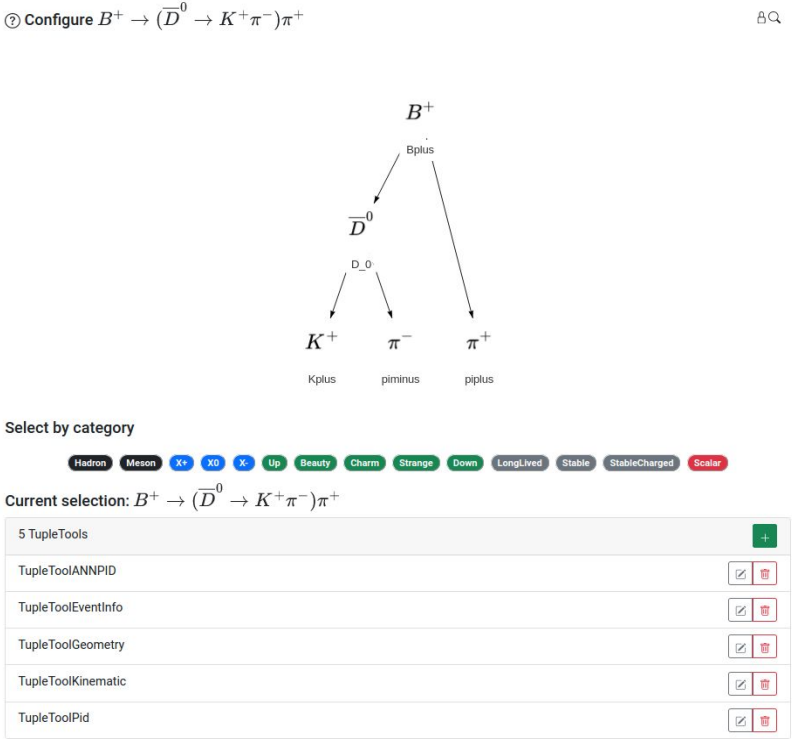
Stripping line selection
- Specifies algorithms applied to identify candidates of the selected physics object

Dataset selection
- Specifies the dataset to run over -- multiple selections can be made



At this stage, the user can initiate configuration of the Ntuple corresponding to the selected physics object(s)

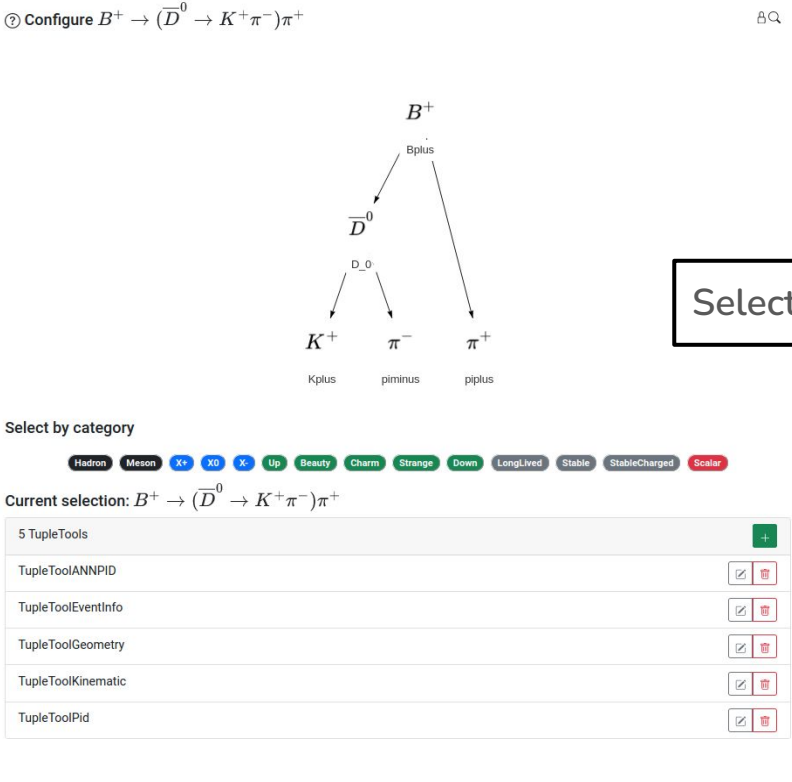
Web Interface: Ntuple Configuration



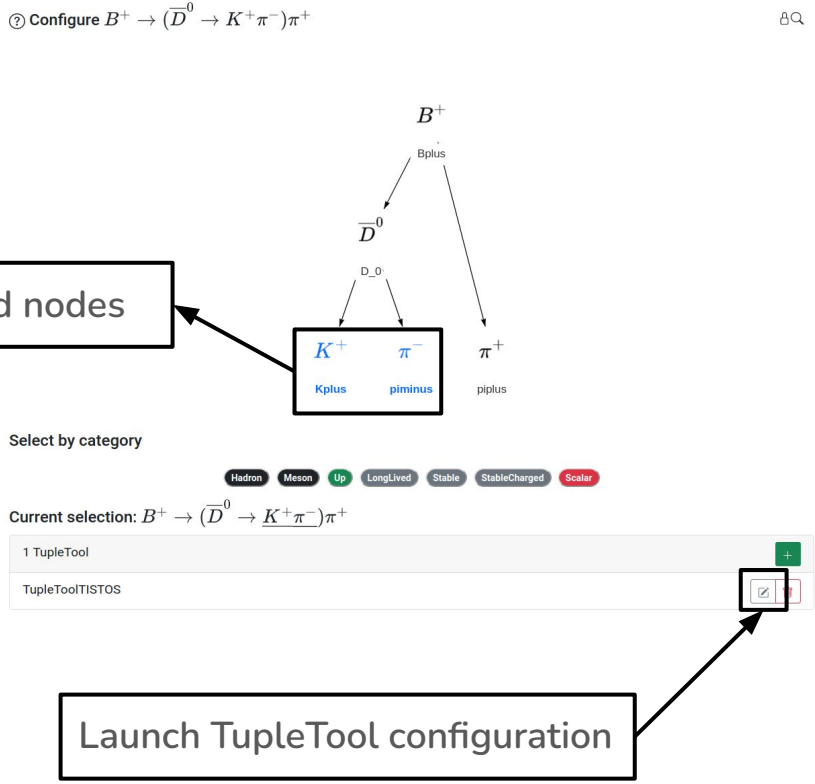
Ntuple configuration via an interactive node tree

- Particles in decay rendered as nodes in tree
- Each node can be configured independently, or in various groupings
  - Labels provided to select nodes by similar categories
- Node configuration proceeds by adding, removing, or configuring ***TupleTools***, which save various physics quantities to the Ntuple
  - Can be performed on entire tree, single node, or selection of nodes
- The entire node tree includes 5 standard TupleTools for LHCb analysis by default

Web Interface: Ntuple Configuration



Selected nodes

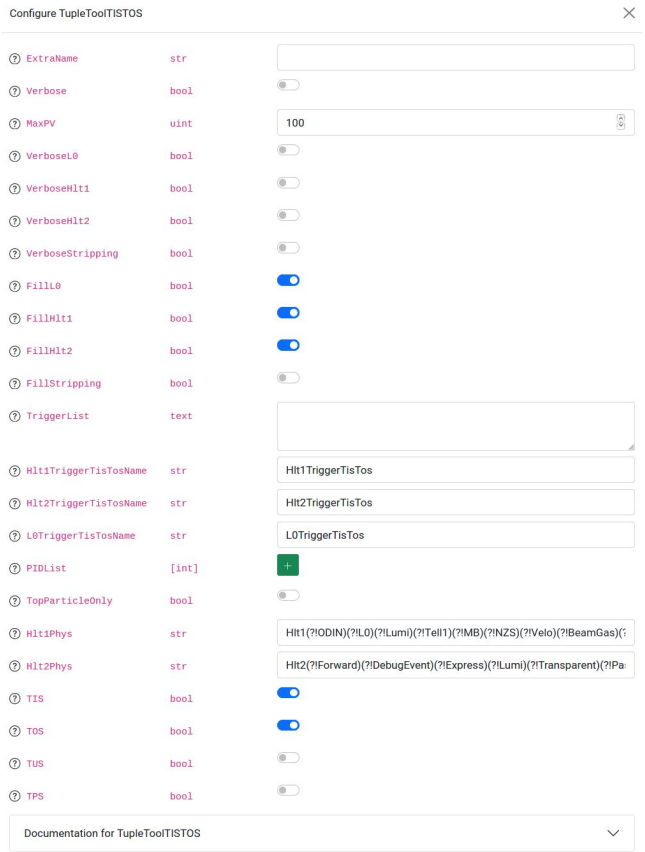


Launch TupleTool configuration

Web Interface: TupleTool Configuration Example

Example of TupleTool configuration interface for TupleToolTISTOS (**T**rigger **I**ndependent of **S**ignal/ **T**rigger **o**n **S**ignal)

- Configurable names, data types, and user input fields are included
- Mouseover tooltips and links to documentation are included for guidance
  - This includes LHCb Doxygen documentation
- Each TupleTool has specific configurables
  - For many tools, the standard configuration is perfectly fine
  - Only certain tools (e.g. related to the trigger) need specific configurations, to be specified in the documentation



Ntuple Configuration Output Example

```yaml
inputs:
  - /Event/BhadronCompleteEvent/Phys/B2D0PiD2HHBeauty2CharmLine/
      Particles
descriptorTemplate: ${Bplus}[B+ -> ${D_0}(D~0 -> ${Kplus}K+ ${piminus}pi
   -)${piplus}pi+]CC
tools:
  - TupleToolKinematic:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      Transporter: ParticleTransporter:PUBLIC
  - TupleToolPid:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
  - TupleToolANNPID:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      ANNPIDTunes:
        - MC12TuneV2
        - MC12TuneV3
        - MC12TuneV4
        - MC15TuneV1
      PIDTypes:
        - Electron
        - Muon
        - Pion
        - Kaon
        - Proton
        - Ghost
  - TupleToolGeometry:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
      RefitPVs: false
      PVReFitter: LoKi::PVReFitter:PUBLIC
      FillMultiPV: false
  - TupleToolEventInfo:
      ExtraName: ''
      Verbose: false
      MaxPV: 100
branches:
  Bplus:
    particle: B+
    tools: []
  D_0:
    particle: D~0
    tools: []
  Kplus:
    particle: K+
    tools: []
  piminus:
    particle: pi-
    tools: []
  piplus:
    particle: pi+
    tools: []
groups:
  Kplus,piminus:
    particles:
      - K+
      - pi-
    tools:
      - TupleToolTISTOS:
          ExtraName: ''
          Verbose: false
          MaxPV: 100
          VerboseL0: false
          VerboseHlt1: false
          VerboseHlt2: false
          VerboseStripping: false
          FillL0: true
          FillHlt1: true
          FillHlt2: true
          FillStripping: false
          TriggerList: []
          Hlt1TriggerTisTosName: Hlt1TriggerTisTos
          Hlt2TriggerTisTosName: Hlt2TriggerTisTos
          L0TriggerTisTosName: L0TriggerTisTos
          PIDList: []
          TopParticleOnly: false
          Hlt1Phys: >-
            Hlt1(?!ODIN)(?!L0)(?!Lumi)(?!Tell1)(?!MB)(?!NZS)(?!Velo)(?!
              BeamGas)(?!Incident).*Decision
          Hlt2Phys: >-
            Hlt2(?!Forward)(?!DebugEvent)(?!Express)(?!Lumi)(?!
              Transparent)(?!PassThrough).*Decision
          TIS: true
          TOS: true
          TUS: false
          TPS: false
name: DecayTreeTuple/Btree
```

Output in pure data structure (YAML) format

- Ntuple configuration output shown based on selections outlined in the previous slides
- An additional yaml file is generated to specify the dataset location and organize the request for production jobs (not shown here)

The YAML files are parsed internally to generate the necessary python options files for the Ntuple production jobs

We are just getting started

Challenges with LHCb Open Data release

Things to improve or add

- Calibration samples + tools
- Documentation on available MC samples
- Analysis example + runtime environment

- MDST and DST are very specialized data formats
  - Ntuple Wizard will write plain ROOT ntuples
  - Ntuple Wizard provides much clearer representation of the content of the data
- Integrate Ntuple Wizard with Open Data Portal (activity ongoing)

- Training for outside users (see CMS Open Data workshops)
- All activity currently severely limited by available people within LHCb

Going beyond level 3 data

Open science and Open data policies:

We should publish ntuples and statistical models to make our results more impactful

5. Research integrity, reuse and reproducibility

CERN is committed to ensuring the integrity of research. In order to facilitate the reuse of its research products, CERN provides infrastructures to accommodate the scale and complexity of its research outputs. Reuse and reproducibility are facilitated by practising comprehensive analysis preservation to capture relevant research objects, such as research data releases with supporting metadata, auxiliary data, linked software, reproducible analysis workflows, documentation, etc.

what to publish depends on individual analysis

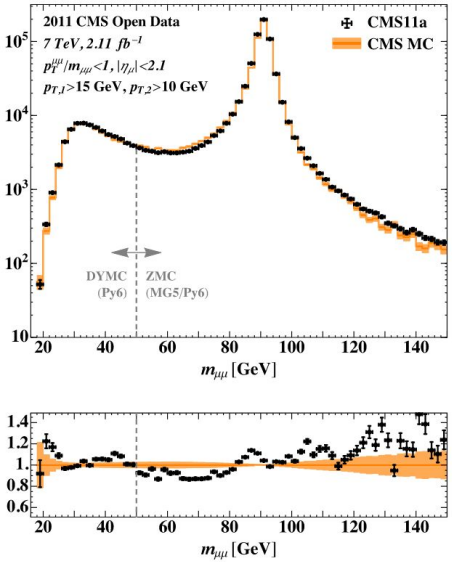
Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their 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.

How is LHC Open Data going to be used?

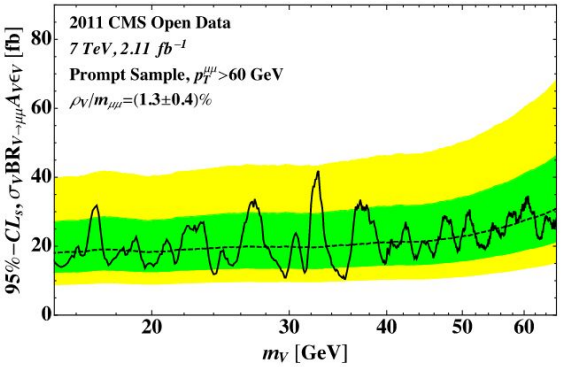
No experience for LHCb data, yet.

But various studies done on CMS open data.
Overview: arXiv:2106.05726

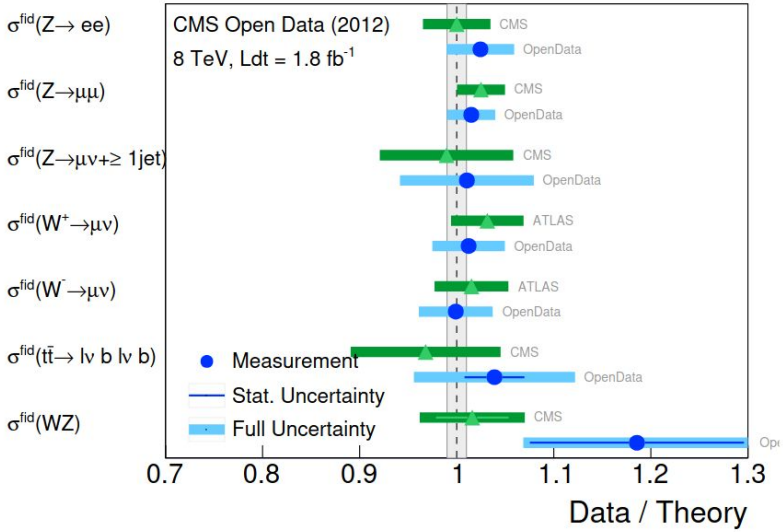
Dimuon spectrum [PRD100(2019)015021]:



Search for narrow dimuon resonances



SM cross section measurements on CMS open data [1907.08197]



Backup

The Open Science Philosophy (at CERN)

Recognize the universal importance of the fundamental scientific knowledge produced at CERN and the key role of openness in the pursuit of CERN organisational mission.

Commits to the advancement of science and wide dissemination of knowledge by adopting practices to make scientific research more open, global, collaborative and responsive to societal changes.

In fulfilment of the collective moral and fiduciary responsibility to member states and the broader global scientific community

Data collected at the LHC is a heritage to humanity.

It has been obtained through collaborative work using public funds.

Therefore, CERN is committed to preserve, curate, steward and share the data with the public.

Open Science Landscape - Recent Trends

- Funding agencies: requests for data management plans
- Publishers: requests for data products allowing to
  - validate / reproduce results
  - reuse data for further studies

Science Community: “**Data is not enough**”:

- Papers with code https://paperswithcode.com/
- Interactive publications
- Federated infrastructures and computing/science portals (e.g. NFDI)
- Not a new realization (see e.g. DPHEP study group 2013 status report) but technology (esp cloud computing, containerization) has made progress!
- Development driven especially through bioinformatics and machine learning / AI community

Open data policy - preamble

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.

Open data policy process and documents

Working group with two representatives from each LHC experiment + CERN IT and Management.

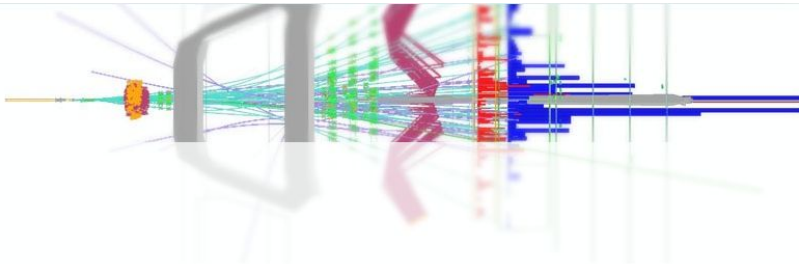
3/4 year process for drafting and ratification.

Policy documents ratified by the collaboration boards of the 4 big experiments.

Two documents prepared.

- Policy
  - short, concise document
  - public
  - specifies commitments in generic form
  - universal for all LHC experiments
- Implementation document
  - longer, technical, “living” document
  - Not public, but can be shown to funding agencies
  - contains details on data volumes, formats and release schedules
  - Individual sections for each experiment, accommodates differences in release strategies

The LHCb Public results



LHCb publications

[to restricted-access page]

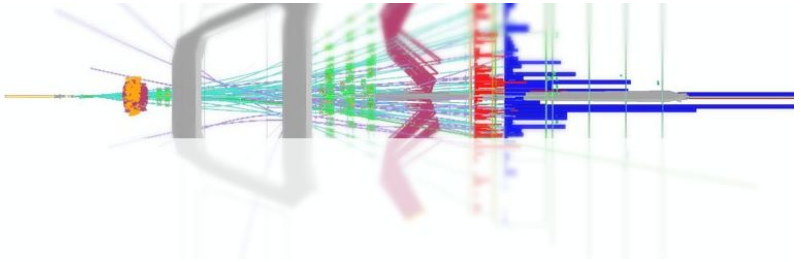
PUBLICATIONS PER WORKING GROUP

- B DECAYS TO CHARMONIUM
- B DECAYS TO OPEN CHARM
- CHARMLESS b-HADRON DECAYS
- b-HADRONS AND QUARKONIA
- CHARM PHYSICS
- FLAVOUR TAGGING
- LUMINOSITY
- QCD, ELECTROWEAK AND EXOTICA
- RARE DECAYS
- SEMILEPTONIC B DECAYS
- DETECTOR PERFORMANCE
- IONS AND FIXED TARGET

List of papers (Total of 655 papers and 50538 citations)

TITLE	DOCUMENT NUMBER	JOURNAL	SUBMITTED ON	CITED
Measurement of Υ production in pp collisions at $\sqrt{s} = 5$ TeV	PAPER-2022-036 arXiv:2212.12664 [PDF]	JHEP	24 Dec 2022	
First observation and branching fraction measurement of the $\Lambda_b^0 \to D_s^- p$ decay	PAPER-2022-038 arXiv:2212.12574 [PDF]	JHEP	23 Dec 2022	
Search for rare decays of D^0 mesons into two muons	PAPER-2022-029 arXiv:2212.11203 [PDF]	PRL	21 Dec 2022	
Measurement of lepton universality parameters in $B^+ \to K^+\ell^+\ell^-$ and $B^0 \to K^{*0}\ell^+\ell^-$ decays	PAPER-2022-045 arXiv:2212.09153 [PDF]	PRD	18 Dec 2022	10
Test of lepton universality in $b \to s\ell^+\ell^-$ decays	PAPER-2022-046 arXiv:2212.09152 [PDF]	PRL	18 Dec 2022	9
Search for the rare decays $W^+ \to D_s^+\gamma$ and $Z \to D^0\gamma$ at LHCb	PAPER-2022-033 arXiv:2212.07120 [PDF]	Chin. Phys. C	14 Dec 2022	
Search for $K^0_{S(L)} \to \mu^+\mu^-\mu^+\mu^-$ decays at LHCb	PAPER-2022-035 arXiv:2212.04977 [PDF]	PRD	09 Dec 2022	
Amplitude analysis of $B^0 \to \overline{D}^0 D_s^+\pi^-$ and $B^+ \to D^- D_s^+\pi^+$ decays	PAPER-2022-027 arXiv:2212.02717 [PDF]	PRD	06 Dec 2022	
First observation of a doubly charged tetraquark and its neutral partner	PAPER-2022-026 arXiv:2212.02716 [PDF]	PRL	06 Dec 2022	
J/ψ and D^0 production in $\sqrt{s_{NN}}$ = 68.5 GeV PbNe collisions	PAPER-2022-011 arXiv:2211.11652 [PDF]	EPJC	21 Nov 2022	
Charmonium production in pNe collisions at $\sqrt{s_{NN}} = 68.5$ GeV	PAPER-2022-014 arXiv:2211.11645 [PDF]	EPJC	21 Nov 2022	
Open charm production and asymmetry in pNe collisions at $\sqrt{s_{NN}} = 68.5$ GeV	PAPER-2022-015 arXiv:2211.11633 [PDF]	EPJC	21 Nov 2022	
Searches for the rare hadronic decays $B^0 \to p\bar{p}p\bar{p}$ and $B_s^0 \to p\bar{p}p\bar{p}$	PAPER-2022-032 arXiv:2211.08847 [PDF]	PRL	16 Nov 2022	

The LHCb Public results



Search for rare decays of D^0 mesons into two muons

[to restricted-access page]

INFORMATION

LHCB-PAPER-2022-029

CERN-EP-2022-273

ARXIV:2212.11203 [PDF]

(SUBMITTED ON 21 DEC 2022)

PRL

INSPIRE 2616985

TOOLS

GET BIBTEX

Abstract

A search for the very rare $D^0 \to \mu^+\mu^-$ decay is performed using data collected by the LHCb experiment in proton-proton collisions at $\sqrt{s} = 7$, 8 and 13TeV, corresponding to an integrated luminosity of 9fb^{-1}. The search is optimised for D^0 mesons from $D^{*+} \to D^0\pi^+$ decays but is also sensitive to D^0 mesons from other sources. No evidence for an excess of events over the expected background is observed. An upper limit on the branching fraction of this decay is set at $\mathcal{B}(D^0 \to \mu^+\mu^-) < 3.1 \times 10^{-9}$ at a 90% CL. This represents the world's most stringent limit, constraining models of physics beyond the Standard Model.

Figures and captions

Distributions of Δm for (left) $D^0 \to K^-\pi^+$ and (right) $D^0 \to \pi^+\pi^-$ normalisation channels candidates for (top) Run 1 and (bottom) Run 2 data. The distributions are superimposed with the fit.

Fig1a.pdf [32 KiB]
HiDef png [192 KiB]
Thumbnail [154 KiB]

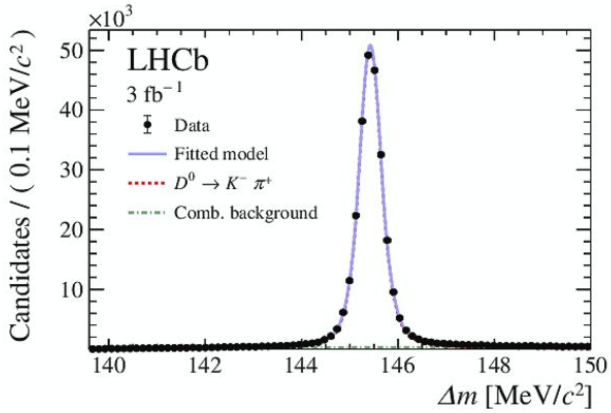


Fig1b.pdf [32 KiB]
HiDef png [195 KiB]
Thumbnail [154 KiB]

Open data curation