Designing and implementing the LHC Open Data policy at CERN

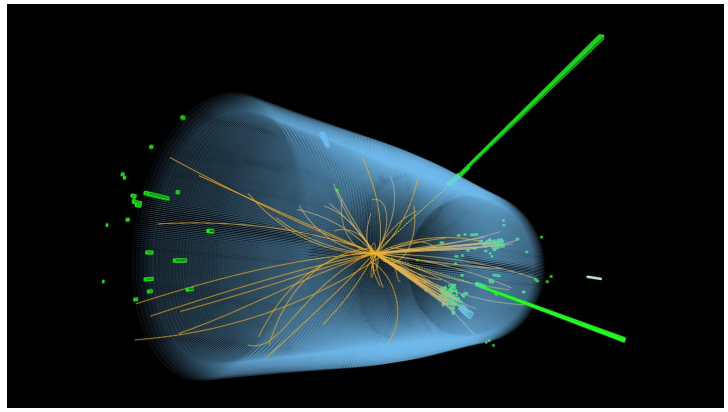
Jamie Boyd (CERN)

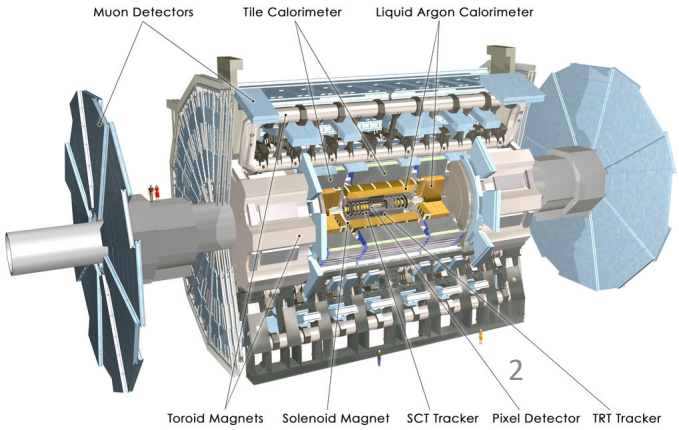
14 December 2023

Policy developed along with the LHC experiments:

Context

- LHC physics
  - Study high energy frontier (discovery of Higgs boson, search for new particles/forces etc…)
  - Complex detectors (millions of channels), large data volumes (O(50) PB/year raw data) etc..
- LHC community
  - 4 Large international collaborations (up to 3000 scientists / collaboration)
  - Collaboration lifetimes several decades
  - CERN as host laboratory
  - Collaborations have their own governance
- Increasing importance of Open Data
  - European Commission (relevant for funding applications)
  - European Stratgey for Particle Physics (2020 update)
- CERN management mandated working group to explore a common open data policy across the collaborations
  - To be formally endorsed by the Collaborations
- Expected community to use LHC Open Data
  - Professional physicists
  - Non-physics scientists (e.g. computer scientists: machine learning, reconstruction algorithms…)
  - Interested non-scientists





Challenges

- Concerns from the collaborations:
  - Ownership
    - Scientists who worked to design (R&D), build, operate the detectors unwilling to lose “ownership” of their data
  - Effort
    - Concern from experiment management that could lose effort to operate the experiment if people can analyze the open data without contributing to the experiment
      - Open Data policy of one LHC collaboration can effect other LHC collaborations => push for common policy!
  - Scientific rigour
    - Worry about lack of scientific rigour in analysis of open data (spurious claims)
  - Resources
    - Required resources within experiments (preparing open data datasets, documentation, storage space etc…), and from CERN side (person power and computing resources)

LHC data levels

- HEP Data Preservation (DPHEP) study group, divided particle physics data into 4 levels:
  - Level 1: scientific papers and associated auxiliary data
  - Level 2: data tailored for outreach and education purposes
  - Level 3: output of data reconstruction. The input for physics analysis.
  - Level 4: the raw data from the experiment

Small

Large

- Prior to the OD policy, all collaborations released Level 1, 2 data, and all agree that level 4 data is not useful for external bodies
  - Nearly all discussion was on Level 3 data

Main points in the policy

- **Level 1 data:**
  - Continue to release, including as much auxiliary data as possible to allow re-interpretation of the results (HepData database)
- **Level 2 data:**
  - Continue to release in appropriate formats/schedule
- **Level 3 data:**
  - Next slide
- **Level 4 data:**
  - Not useful, will not be released

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.

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 exploration of these

CERN-OPEN-2020-013
20/11/2020

[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.

1

After ~1 year of discussions the policy was endorsed by the large LHC experiments in late 2020.
Policy document: https://cds.cern.ch/record/2745133

Main points in the policy

- Level 3 data:
  - Release data within 5 years after end of running period
    - *Latency key to counter resistance from within the collaborations*
    - Collaboration can withhold releasing data in special circumstances (unfinished high profile analysis ongoing)
  - Exact format determined by collaboration
    - Same format as used internally in the collaboration for physics
  - Also release analysis software and simulated data samples
    - Needed to allow *meaningful scientific study of the data*
  - Documentation / support offered on best effort basis
  - Data released via CERNs OpenData portal
    - Storage media supplied by CERN
      (may not be long term solution, but for first 5 year period)
- Open Data policy important for preparing experiments for long term data preservation needs
  - Many aspects of Open Data relevant for Data Presevation

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.

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 exploration of these

CERN-OPEN-2020-013
20/11/2020

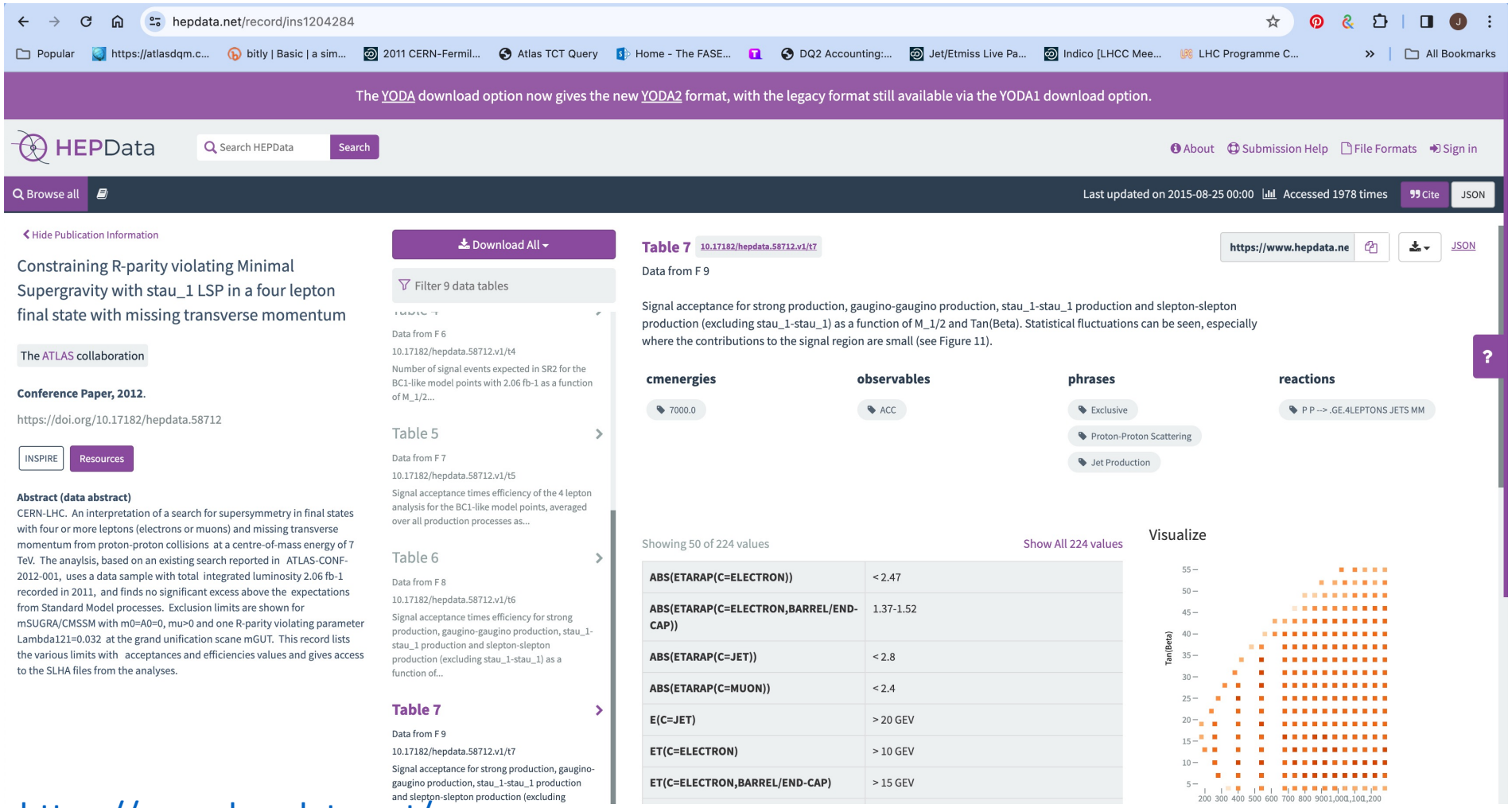
[1] European Strategy Group (2020), '2020 Update of the European Strategy for Particle Physics'.
[2] FAIR Guiding Principles for scientific data management and stewardship. Available at: https://www.go-fair.org/fair-principles/.
[3] Data management plans are defined by the LHC experiments to address 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.

1

After ~1 year of discussions the policy was endorsed by the large LHC experiments in late 2020.
Policy document: https://cds.cern.ch/record/2745133

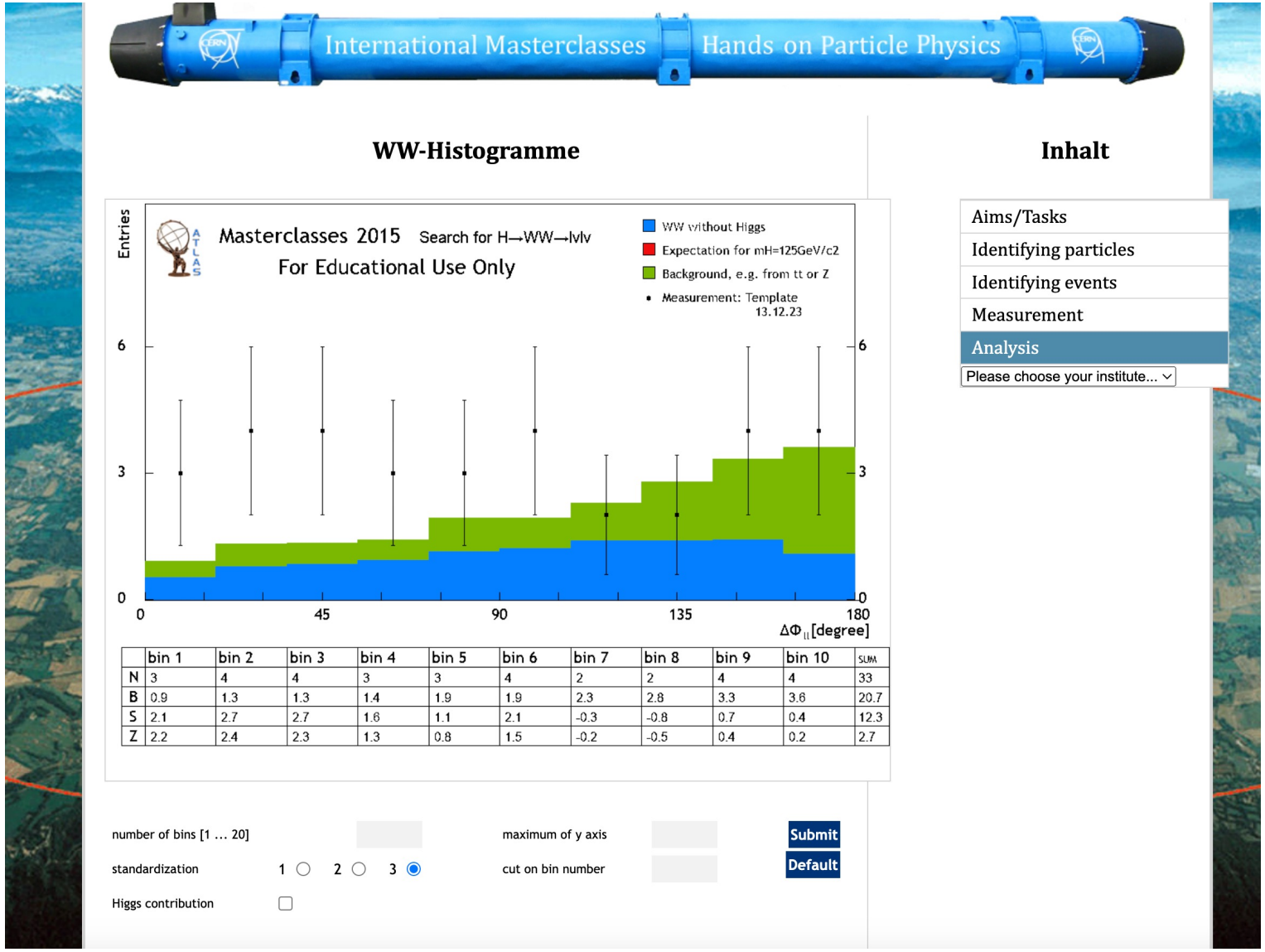
Level 1 Open Data

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



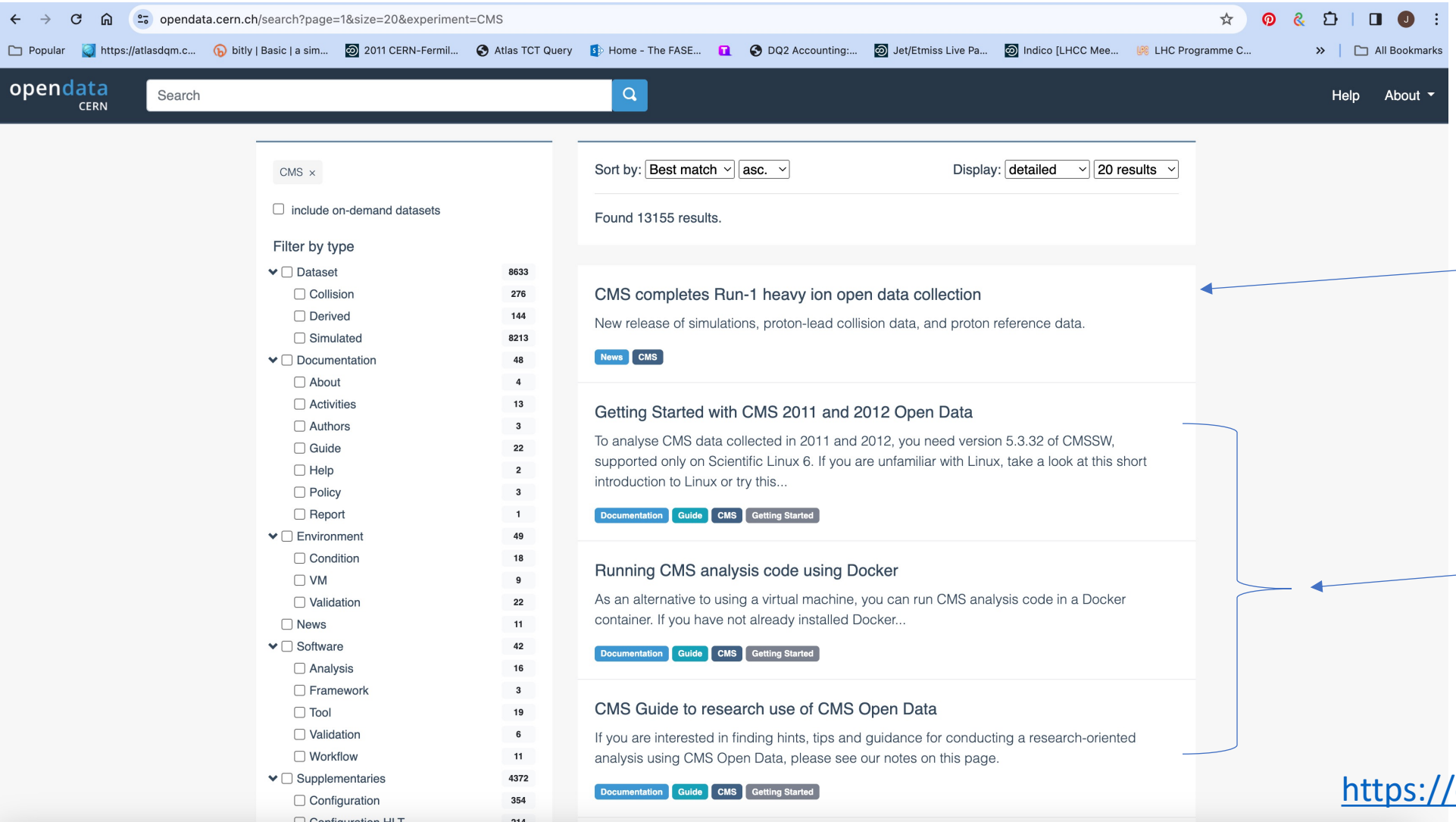
	bin 1	bin 2	bin 3	bin 4	bin 5	bin 6	bin 7	bin 8	bin 9	bin 10	SUM
N	3	4	4	3	3	4	2	2	4	4	33
B	0.9	1.3	1.3	1.4	1.9	1.9	2.3	2.8	3.3	3.6	20.7
S	2.1	2.7	2.7	1.6	1.1	2.1	-0.3	-0.8	0.7	0.4	12.3
Z	2.2	2.4	2.3	1.3	0.8	1.5	-0.2	-0.5	0.4	0.2	2.7

Example use of real experimental data for education and outreach. In this ATLAS masterclass, school students can apply selections to real ATLAS 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://atlas.physicsmasterclasses.org/

Level 3 Open Data

The CERN Open Data Portal is where the Level 3 Open Data is stored and can be accessed from.
Typically a few PB of Level 3 Open data will be released through the portal per year.



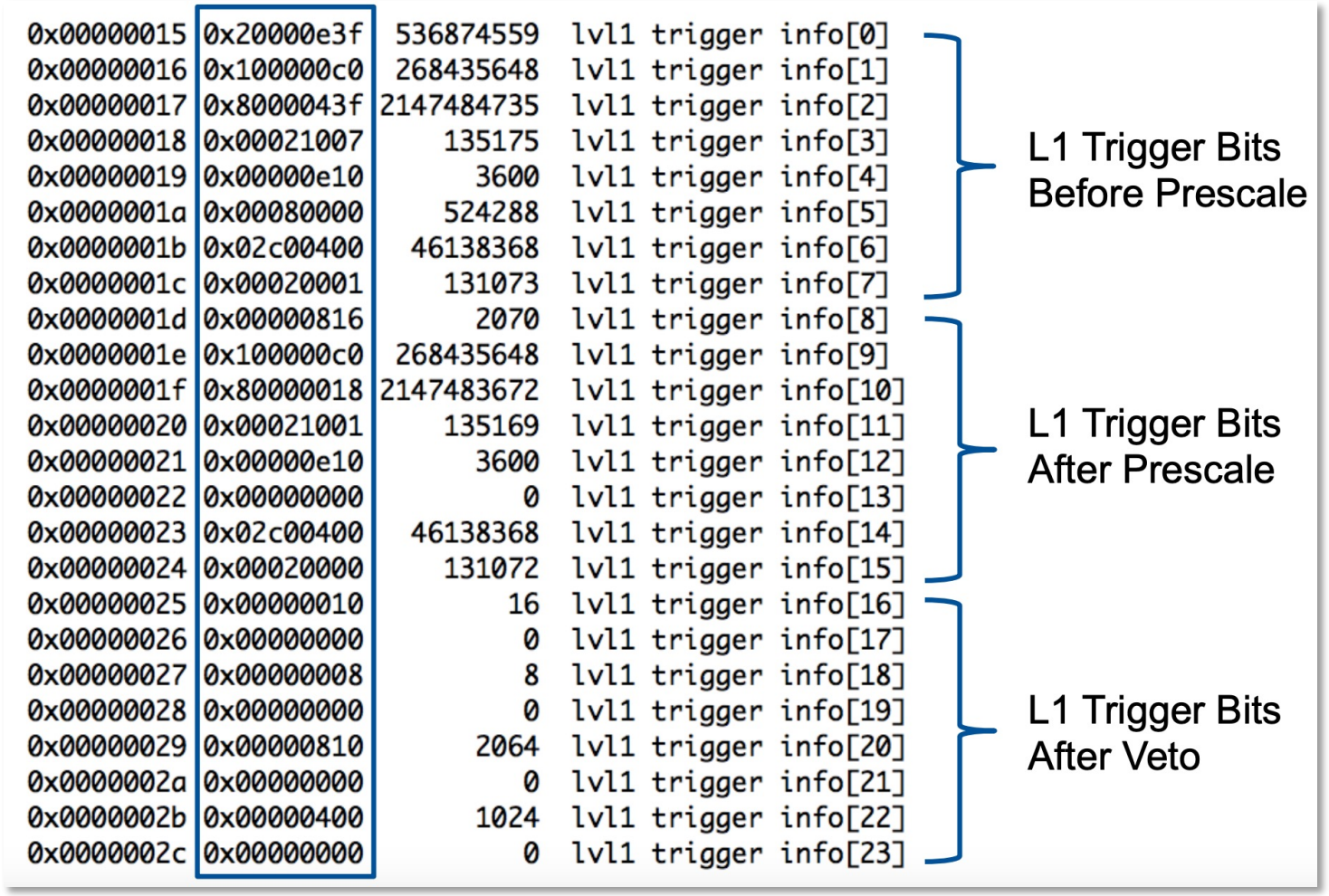
A 300TB dataset

Useful information/ documentation

https://opendata.cern.ch/

Level 4 ~~Open~~ Data

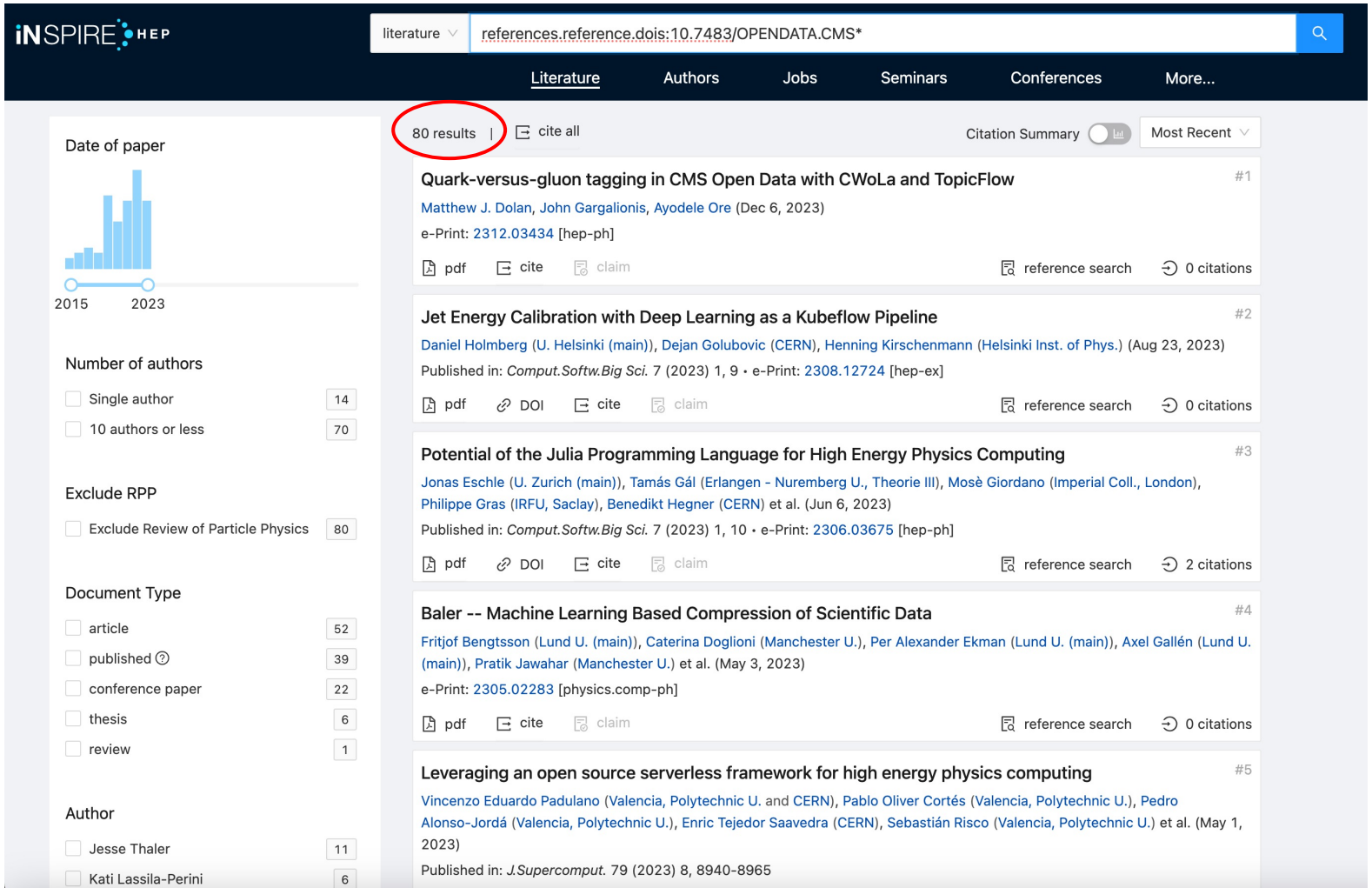
A tiny fraction of 1 RAW event from the ATLAS experiment:



The LHC experiments are incredibly complex, with millions of readout channels covering many detector technologies.
The programs to reconstruct the raw data are hugely complex and need a lot of expertize in the experiment to be able to use in a useful way.
It is therefore not realistic for people from outside of the experiment to be able to extract meaningful results from the raw data.

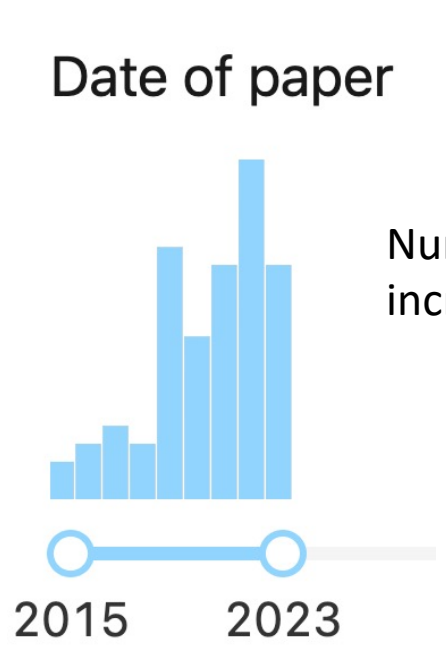
Level 3 Open Data Use

The policy discussed in this talk started at the beginning of 2021, and with up to 5 years latency.
Therefore, not so much data has been released through the policy to date.
However, the CMS experiment has been releasing Open Data for several years.



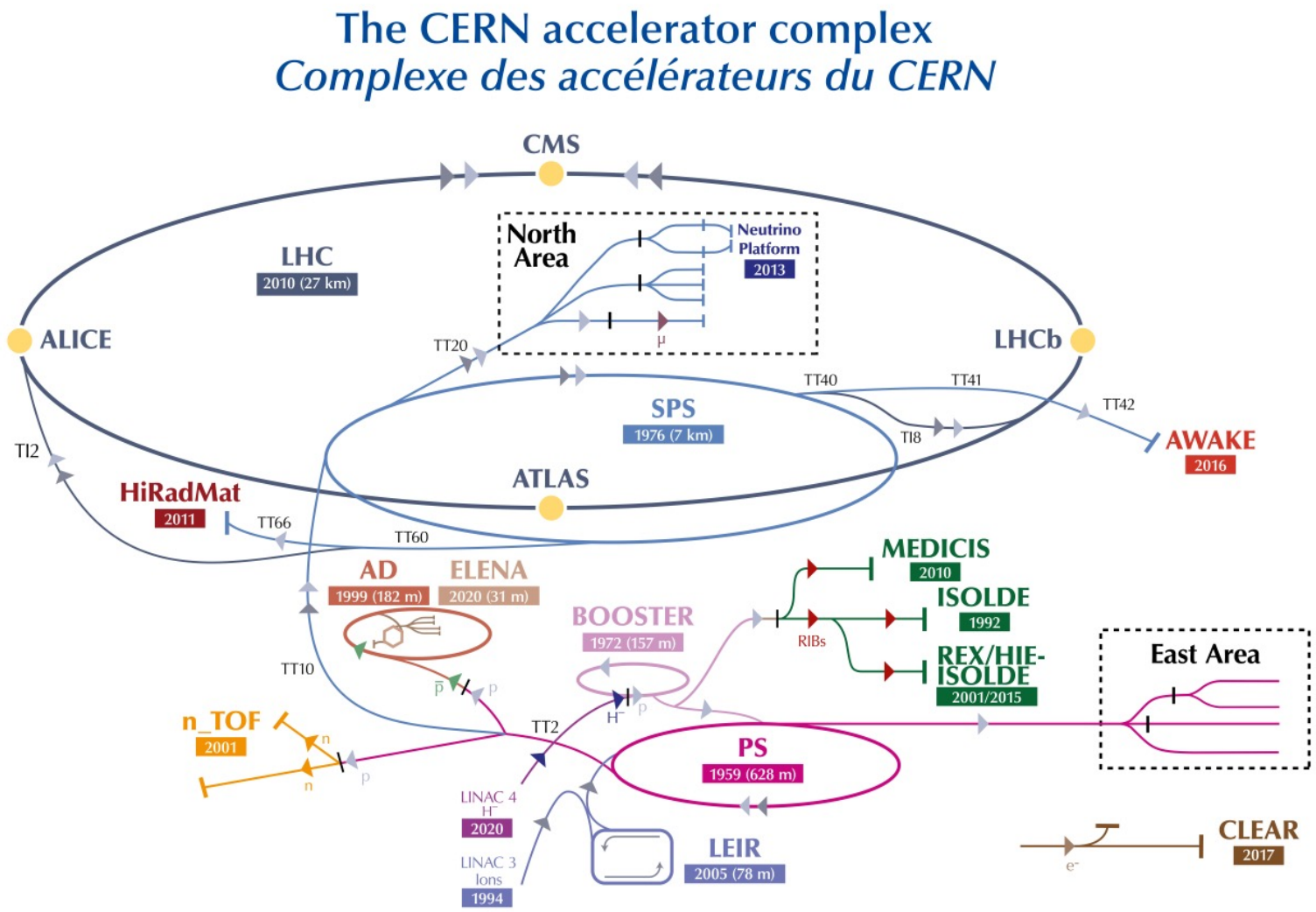
80 papers have been released using that data.
Mostly covering:

- Pure physics research
- Using Machine Learning techniques to improve event reconstruction



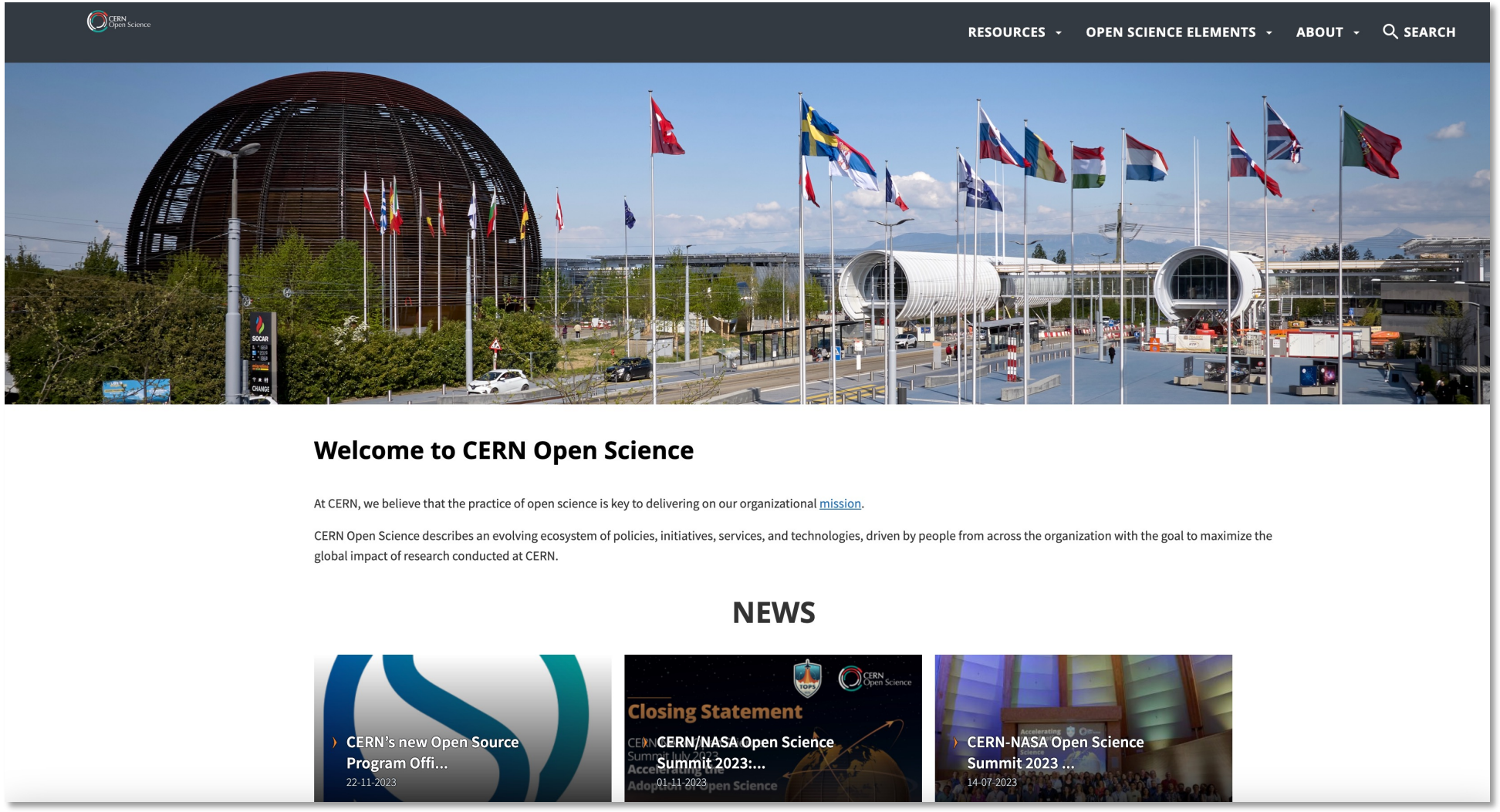
Number of papers increasing per year

Future Plans



- CERN has a diverse accelerator complex feeding many experiments, not just the large-LHC experiments discussed so far.
- CERN management has asked us to start the process of getting the smaller CERN experiments to sign up to the Open Data policy.
- So far the initial response to this has been generally positive, but there are several more steps to go through.
- In my experience the smaller experiments are less worried about issues related to "*effort*" and "*ownership*" than the large LHC experiments, but are more worried about "*resources*" since they have less people in the experiment.

Open Science at CERN



Since the Open Data policy was released CERN has produced its broader Open Sceience policy (released in 2022).
This covers the following areas:

- Open Access
- Open Data +
- Open Source Software
- Open Source Hardware
- Research Integrity
- Open Infrastructure
- Research Assessment
- Training & Outreach
- Citizen Science

Open Science Policy: https://cds.cern.ch/record/2856044/

Summary

- The CERN LHC Open Data policy was released at the end of 2020
- It was crafted to find the balance between the wish to open the data, with the constraints/concerns of the experiments
  - We believe a good balance was found in the final policy document
- Following this, Open Data has been released and is being used
- We are now trying to enlarge the policy to cover all experiments at CERN
  - Smaller experiments have different challenges which need to be addressed
- The effort is now embedded in the recent CERN Open Science effort