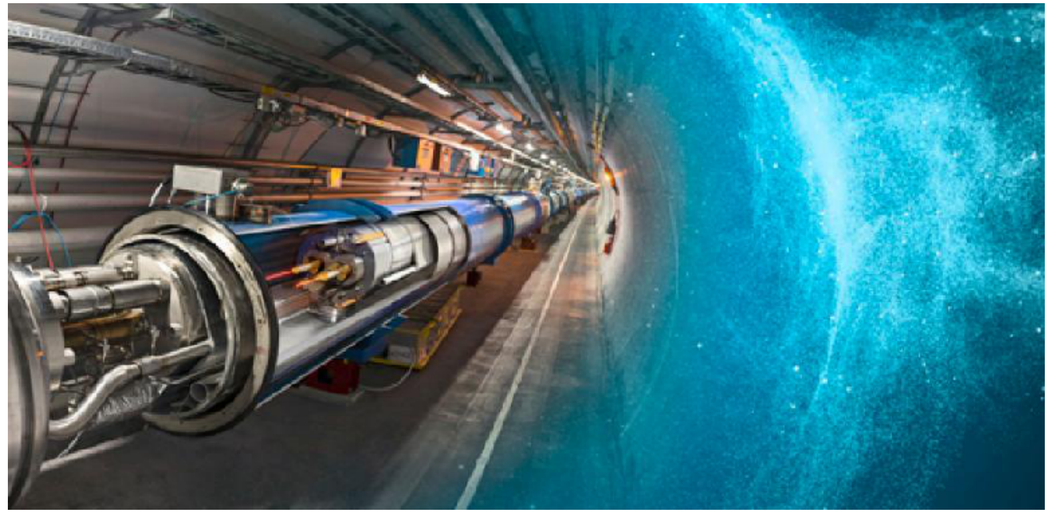
Open Access to Data

Eckhard Elsen

former Director Research and Computing



Unveiling hidden Physics Beyond the Standard Model at the LHC, March 1-3, 2021

Public Research is funded by the tax payer

and hence

data should be available to everyone
(at least to each tax payer)

Benefits of Open Data

- More scientists engage and contribute to the interpretation of the data
  - multiple disciplines
  - foster broader view
  - tools could be re-used
  - faster analysis feedback

and some downsides…

- there may be false scientific claims because of incomplete understanding of the data
  - who will be blamed?
  - who is paying for the effort got correcting the damage?
- Ingenuity of the instrument builder may not be properly recognised
- The experimental exploitation may be incomplete because of lack of familiarity with the tools

data should be available to everyone

Yes, true. But wait a moment…

Tax payer’s second thought…

- „National“ pride
  - Can I draw an advantage from the research conducted to help my economy?
- Partition between **public** and **private** research
  - Private companies would like to profit from publicly funded research (financially and intellectually) but would also want to retain exclusive access to the results
  - Political / societal decision what to support and fund

Fundamental Research from the Tax Payer's View

- Fundamental physics typically too far from immediate application and direct commercial exploitation
- There is a huge interest in understanding the working of the world or the universe at a principal and fundamental level
  - evidenced by the huge interest of students and people at large
  - need and willingness to fund large facilities to increase knowledge

Makes it easier to release data

Researcher's Point of View

- Progress derives from
  - theoretical and instrumental progress
  - competition between different approaches to eventually arrive at the best tool for exploration
  - and often relentless dedication of a team of scientists to construct and optimise apparatus
    - for large apparatus the times for preparation may be decade(s): LIGO, LHC Experiments, Planck

Solution 1: Separation of Construction and Analysis

- Example Astroparticle physics
  - Team(s) of instrument builders develop and operate the next generation tool
    - COBE, WMAP, Planck, …
    - Fermi Gamma-ray Space Telescope (FGST)
    - …
  - Results become publicly available after rather short time

The interpretation of the data is fairly straightforward

Solution 2: Integration of Construction and Analysis

- LHC physics
  - Team(s) of instrument builders develop and operate the next generation tool over many decades
  - Collaborations are the owners (or custodians) of the recorded data
  - huge effort to record the data
  - mammoth task to trigger, calibrate and analyse the data
  - extremely difficult to explain the detailed features of the recorded data

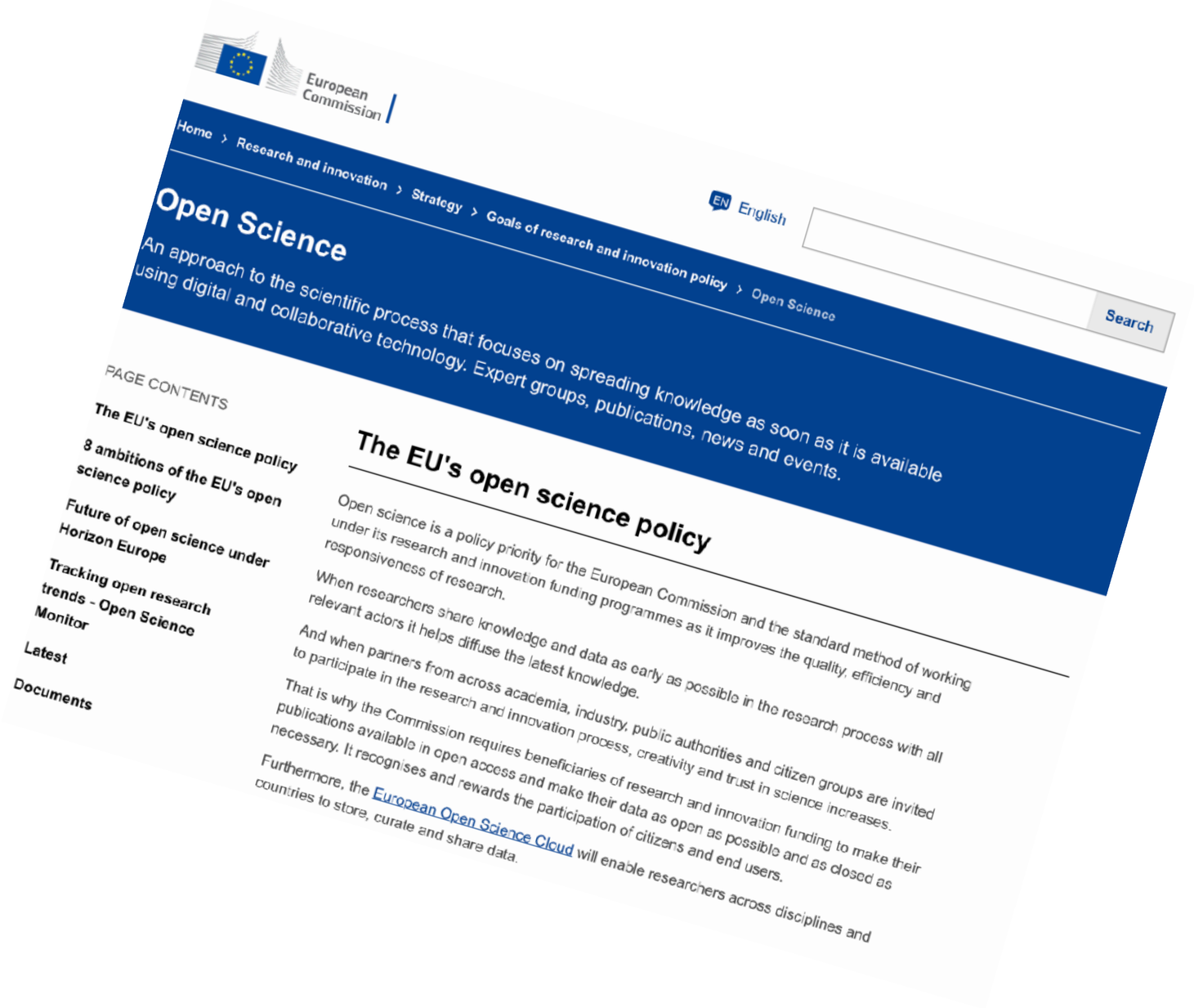
The interpretation of the data is complex

Cost of Open Data for Fundamental Science

- In addition
  - to building the experimental apparatus
  - to recording the data over long periods with dedicated teams
- the data have to be prepared in a fashion that is acceptable to the occasional user
  - aggregated and reduced to the physics information
    (such data will not have the ultimate resolving power)

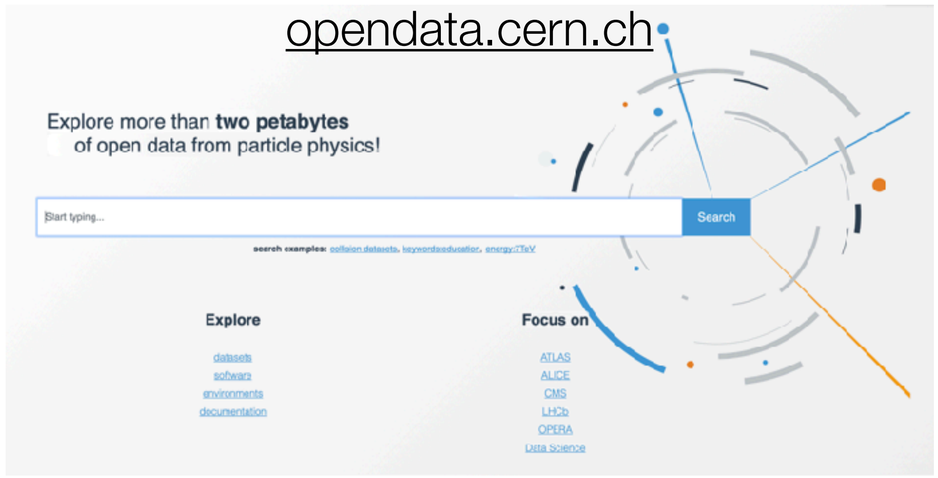
What has happened at the LHC?

Open Data Policy



Open Data Policy Working Group for LHC Experiments

- Working Group distinguished various levels of data
  - Published Results (Level 1)
    - Open Access (for HEP: SCOAP3)
  - Outreach and Education (Level 2)
    - CERN Open Data Portal
  - **Reconstructed Data (Level 3)**
    - Suitable for physics analysis albeit not at ultimate calibration/ resolution
  - Raw Data (Level 4)
    - not really suitable for external consumption



LHC Open Data Policy

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

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

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

[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.

adopted by all 4 LHC Experiments

data may also be provided. CERN experiments will make data of such high level of abstraction available, accessible through the CERN Open Data Portal.[5]

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.

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.

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.

[5] CERN Open Data portal: http://opendata.cern.ch.

data + algorithms

~5 years

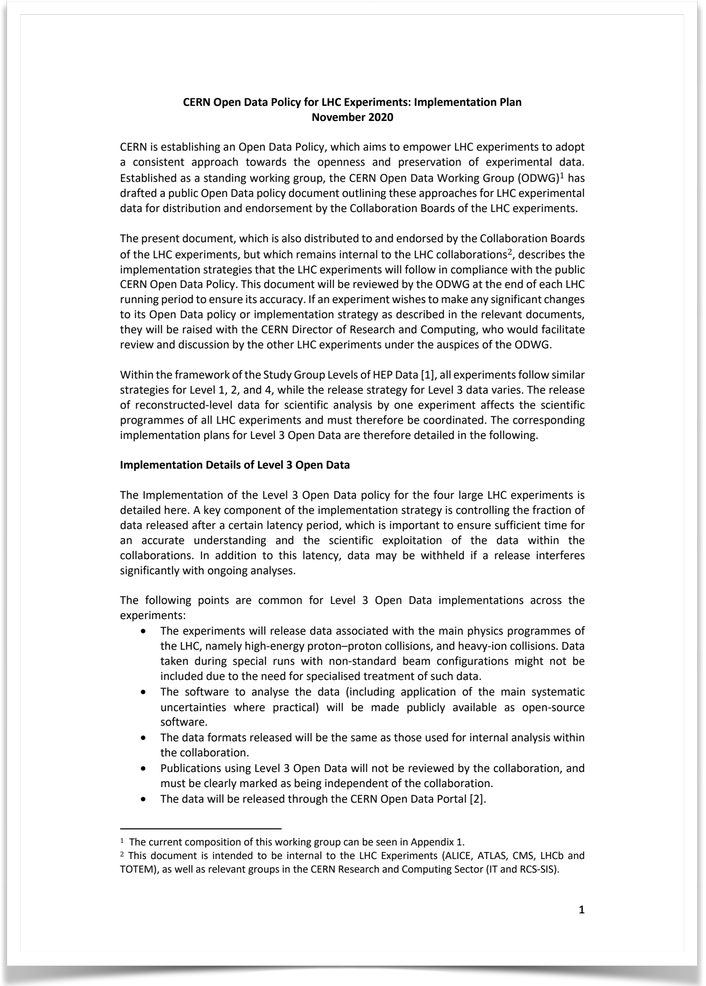
IP

disclaimer

***From the minutes of the SIPB**:*
The Scientific Information Policy Board welcomes the initiative to establish a CERN Open Data Policy, and praises the "ODP Working Group for the LHC experiments" for its preparatory work, for the drafting of the policy document, and for achieving its approval by the experiments. The policy comprises two documents: a public one, outlining the general principles, and an internal document, outlining the implementation of the policy by each experiment. The SIPB strongly supports the CERN principle of openness and its various initiatives towards open science. The new Open Data Policy fully meets this spirit, while meeting the constraints set by the complexity of the LHC data and of their public use. In endorsing the Open Data Policy, the SIPB encourages the CERN management to extend its scope and implementation to cover the non-LHC experiments as well, and remains available to contribute.

Open Data Implementation Document

- Collaborations also agreed on the implementation https://cds.cern.ch/record/2745081 (not public)
  - Latency
  - amount of data
- This document is experiment specific:
  - Heavy Ion data (ALICE) vs Heavy Quark Data (LHCb) vs Searches for ATLAS and CMS



CERN Open Data Policy for LHC Experiments: Implementation Plan
November 2020

CERN is establishing an Open Data Policy, which aims to empower LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Established as a standing working group, the CERN Open Data Working Group (ODWG)[1] has drafted a public Open Data policy document outlining these approaches for LHC experimental data for distribution and endorsement by the Collaboration Boards of the LHC experiments.

The present document, which is also distributed to and endorsed by the Collaboration Boards of the LHC experiments, but which remains internal to the LHC collaborations[2], describes the implementation strategies that the LHC experiments will follow in compliance with the public CERN Open Data Policy. This document will be reviewed by the ODWG at the end of each LHC running period to ensure its accuracy. If an experiment wishes to make any significant changes to its Open Data policy or implementation strategy as described in the relevant documents, they will be raised with the CERN Director of Research and Computing, who would facilitate review and discussion by the other LHC experiments under the auspices of the ODWG.

Within the framework of the Study Group Levels of HEP Data [1], all experiments follow similar strategies for Level 1, 2, and 4, while the release strategy for Level 3 data varies. The release of reconstructed-level data for scientific analysis by one experiment affects the scientific programmes of all LHC experiments and must therefore be coordinated. The corresponding implementation plans for Level 3 Open Data are therefore detailed in the following.

Implementation Details of Level 3 Open Data

The Implementation of the Level 3 Open Data policy for the four large LHC experiments is detailed here. A key component of the implementation strategy is controlling the fraction of data released after a certain latency period, which is important to ensure sufficient time for an accurate understanding and the scientific exploitation of the data within the collaborations. In addition to this latency, data may be withheld if a release interferes significantly with ongoing analyses.

The following points are common for Level 3 Open Data implementations across the experiments:

- The experiments will release data associated with the main physics programmes of the LHC, namely high-energy proton–proton collisions, and heavy-ion collisions. Data taken during special runs with non-standard beam configurations might not be included due to the need for specialised treatment of such data.
- The software to analyse the data (including application of the main systematic uncertainties where practical) will be made publicly available as open-source software.
- The data formats released will be the same as those used for internal analysis within the collaboration.
- Publications using Level 3 Open Data will not be reviewed by the collaboration, and must be clearly marked as being independent of the collaboration.
- The data will be released through the CERN Open Data Portal [2].

[1] The current composition of this working group can be seen in Appendix 1.
[2] This document is intended to be internal to the LHC Experiments (ALICE, ATLAS, CMS, LHCb and TOTEM), as well as relevant groups in the CERN Research and Computing Sector (IT and RCS-SIS).

1

Commitment to Openness

- CERN has been built upon principles of openness and global collaboration amongst researchers. Strong belief in values of openness and collaboration, both for scientific advancement and for peace and understanding

  - Open collaborative approach resulted in major discoveries that no individual institute or country could have achieved. - Open collaboration yields better results than would the sum of the individual contributions

  - The ongoing COVID-19 pandemic shows how much more effective science can and could be by having access to all available data on the disease - yet, need to respect the privacy of the individual.

Topic of 2nd WS

I am pleased to see that some of the data storage platforms at CERN have benefitted SARS-CoV-2 research

Open Source

- The Web was invented at CERN with the goals of optimising the interaction between dispersed teams and of sharing the results of research rapidly and without technical or organisational obstacles

- CERN shared the Web code freely with the world to ensure that everyone could profit from it, and collaboratively improve and build upon it

- CERN contributes strongly to the Open Source community, sharing in-house developed software, reusing existing software wherever possible and contributing back fixes and improvements.

What has been paid by the public should be made available to the public.

Open Access

- Since start of LHC, all related scientific publications are available Open Access.
  - 2014: CERN introduced Open Access policy and provides central resources to make all its scientific articles openly available.
  - 2014: CERN founded and started to host SCOAP3, a global collaboration of 3000 libraries to make all HEP scholarly output openly accessible.
- CERN committed to go beyond access to scientific papers.

Sponsoring Consortium for Open Access Publishing in Particle Physics

- All research artefacts that led to a scientific discovery should be openly accessible and reusable to allow further exploitation of data and reproducibility of results.

Open Data

- CERN piloted the CERN Open Data Portal, sharing many terabytes of LHC data.
  - Applications are test of theoretical models, development of algorithms and exploring new analysis ideas
  - First new physics papers from non-CERN persons resulted from reusing the available data.
- Availability of large structured data sets may benefit other fields (e.g. in machine learning).
- CERN now in the process to conclude a CERN-wide Open Data policy.

Conclusion

- CERN is committed to support Openness for publicly funded developments at all level
  - Open access to research results
  - Open Data
  - Open Software and reuse of the software

… and a corollary

- Open Data naturally leads to data preservation
  - preservation well beyond the life time of the experiments
    - This requires the commitment of the hosting institution and comes at a (hopefully then small) cost

Thank you