DPHEP and Long Term Future Data Preservation

Silvia Amerio (Università di Padova e INFN)

Workshop CCR

Trento - 17 Marzo 2016

DPHEP

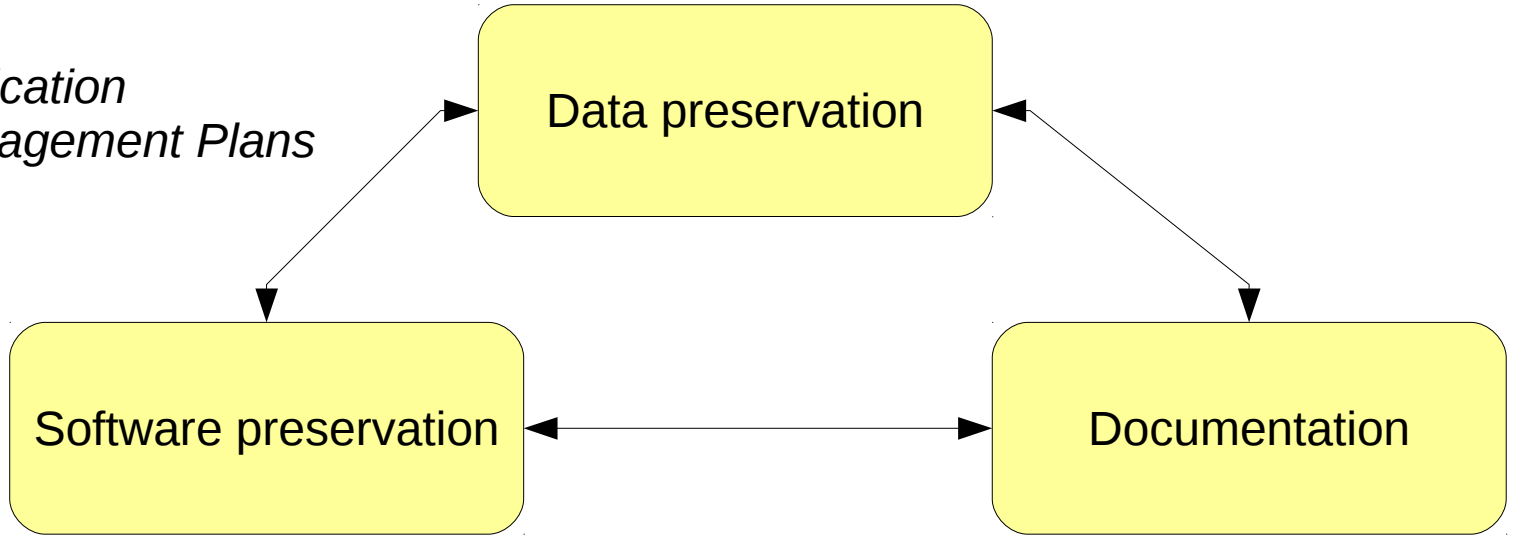
- DPHEP from study group (2009) to collaboration (2012)
- 9 institutes signed the collaboration agreement so far (CERN, DESY, HIP Finland, IHEP, IN2P3, KEK, MPP, IPP and STFC)

The Project, in coordination with the International Committee for Future Accelerators (ICFA), aims at:

1. *Positioning itself as the natural forum for the entire discipline in order to foster discussion, achieve consensus and transfer knowledge in two main areas:*
   a. *Technological challenges in data preservation in HEP,*
   b. *Diverse governance at the collaboration and community level for preserved data,*
2. *Co-ordinate common R&D projects aiming to establish common, discipline-wide preservation tools,*
3. *Harmonize preservation projects across the Partners and liaise with relevant initiatives from other fields,*
4. *Design the long-term organization of sustainable and economic preservation in HEP,*
5. *Outreach within the community and advocacy towards the main stakeholders for the case of preservation in HEP.*

All activities at this link: https://indico.cern.ch/category/4458/

DPHEP activities



Site certification
Data Management Plans

CernVM

Cern Analysis Preservation framework
Open Data portal
DPHEP portal

Data preservation: site certification

- Data are stored in different dedicated computing centers. How can we ensure long term future preservation of data?
- OAIS reference model to set a standard for the activities in preserving a digital archive (What, not How)
- Basic metrics:
  - Are the bits safe?
  - Are the data understandable/usable by the designated community?

ISO 16363 Table of Contents

Three main sections:

Organisational Infrastructure
e.g. The repository shall have a documented history of the changes to its operations, procedures, software, and hardware.

Digital Object Management
e.g. The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.

Infrastructure and Security Risk Management
eg. The repository shall have procedures in place to evaluate when changes are needed to current software.

3 ORGANIZATIONAL INFRASTRUCTURE 3-1

3.1 GOVERNANCE AND ORGANIZATIONAL VIABILITY 3-1
3.2 ORGANIZATIONAL STRUCTURE AND STAFFING 3-3
3.3 PROCEDURAL ACCOUNTABILITY AND PRESERVATION POLICY FRAMEWORK 3-5
3.4 FINANCIAL SUSTAINABILITY 3-10
3.5 CONTRACTS, LICENSES, AND LIABILITIES 3-11

4 DIGITAL OBJECT MANAGEMENT 4-1

4.1 INGEST: ACQUISITION OF CONTENT 4-1
4.2 INGEST: CREATION OF THE AIP 4-6
4.3 PRESERVATION PLANNING 4-16
4.4 AIP PRESERVATION 4-19
4.5 INFORMATION MANAGEMENT 4-23
4.6 ACCESS MANAGEMENT 4-24

5 INFRASTRUCTURE AND SECURITY RISK MANAGEMENT 5-1

5.1 TECHNICAL INFRASTRUCTURE RISK MANAGEMENT 5-1
5.2 SECURITY RISK MANAGEMENT 5-12

Agreed to self-certify Cern T0 by nex DPHEP meeting (October)

Decide what to do with T1 and T2 based on that experience

Data Management Plans

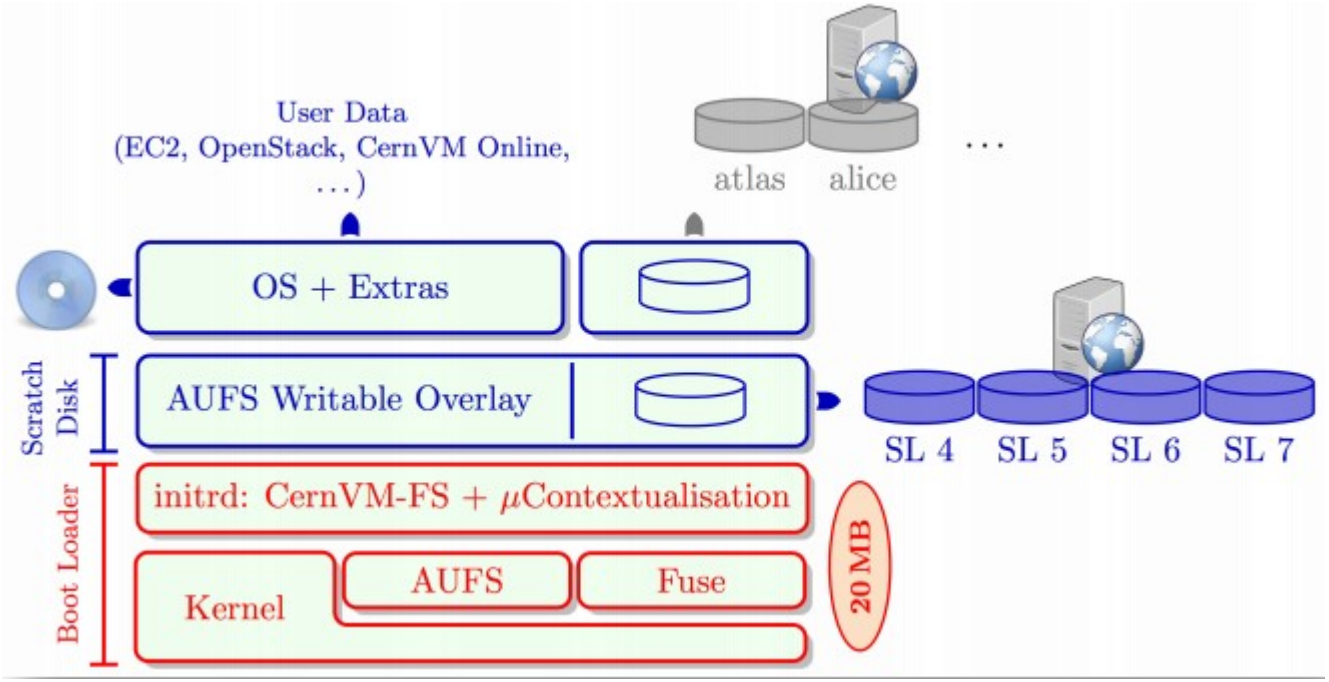
- For new proposals, funding agencies explicitely require DMP to describe how data generated through the course of the proposed research will be shared and preserved.

The DMP should address the points below…

1. Data set reference and name
   - Identifier for the DS to be produced
2. Data set description
   - Description; origin; nature & scale; to whom useful; underpins publication? similar data?
3. Standards and metadata
   - Reference to standards *of the discipline*
4. Data sharing
   - How will it be shared? Embargo periods? Mechanisms for dissemination, s/w and other tools for re-use, access open to restricted to groups, where is repository? Type of repository?
5. Archiving and preservation
   - Description of procedures, how long will it be preserved? End volume? Costs? How will these be covered?

Software preservation: CernVM

CernVM/CernVM File System provide a portable software development and runtime environment for HEP experiments.

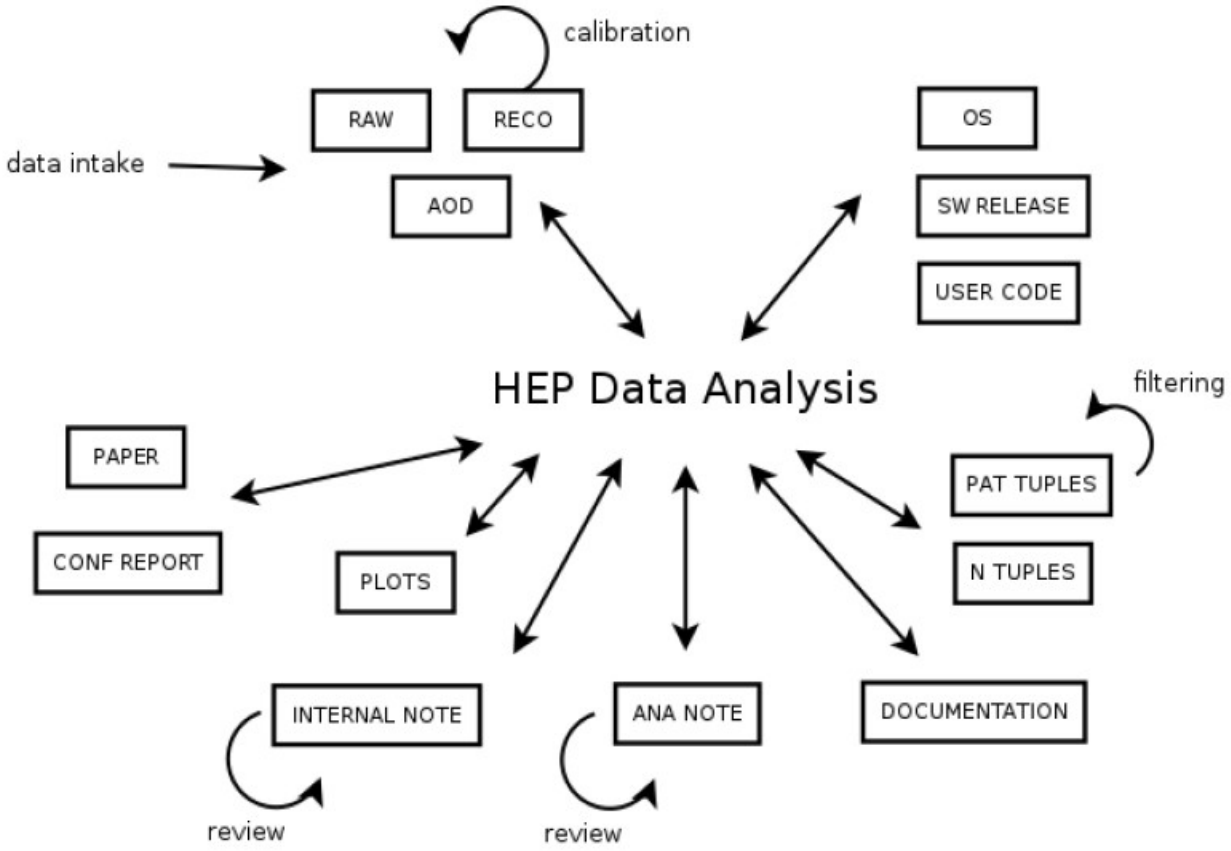


Used in production by all LHC experiments and other scientific collaborations.
Used for open data / data preservation as well:

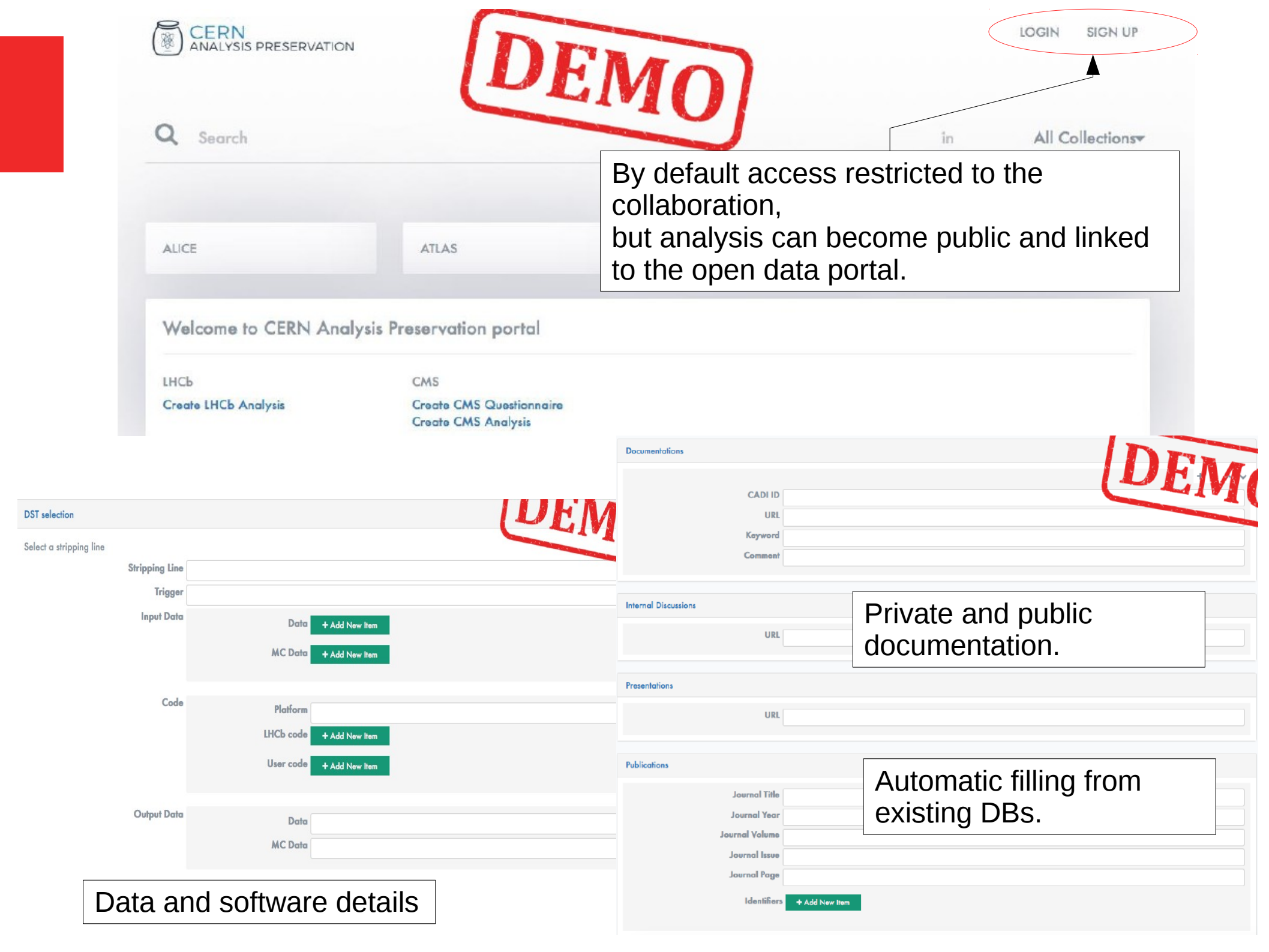
- CMS open data pilot project, to access and analyse CMS 2010 data (SL5);
- Aleph: ALEPH software was installed post-ex on CernVM-FS and made available in a RHEL 4 compatible CernVM on the current CERN OpenStack infrastructure.

Cern analysis preservation framework

Framework to archive and preserve all ingredients of an analysis (data, software, public and private documentation)
Non only an archive --> in the next future, possibility to re-run the analysis (VM images, containers, notebooks).

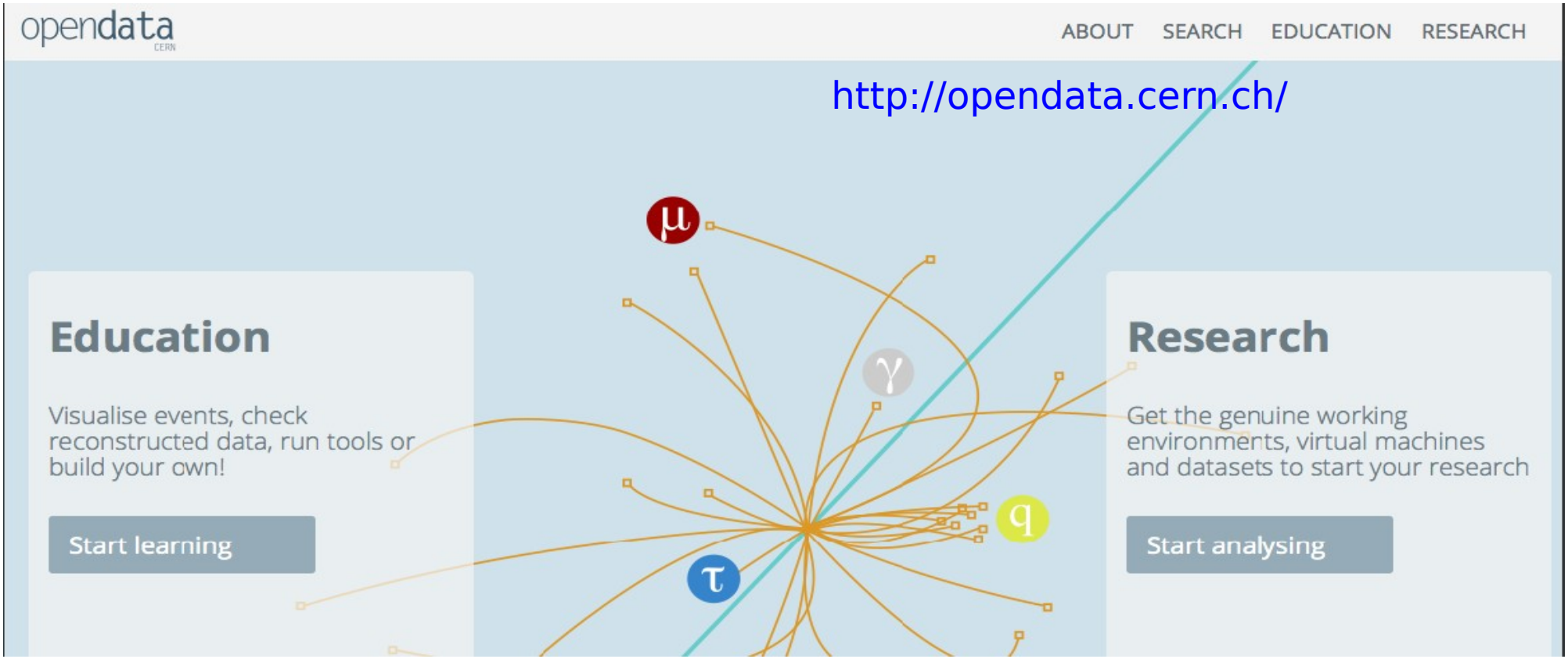


First version ready to be tested by the experiments.


CERN
ANALYSIS PRESERVATION
DEMO
LOGIN
SIGN UP
Search
in
All Collections
ALICE
ATLAS
Welcome to CERN Analysis Preservation portal
LHCb
Create LHCb Analysis
CMS
Create CMS Questionnaire
Create CMS Analysis
By default access restricted to the collaboration,
but analysis can become public and linked to the open data portal.
DST selection
Select a stripping line
Stripping Line
Trigger
Input Data
Data
+ Add New Item
MC Data
+ Add New Item
Code
Platform
LHCb code
+ Add New Item
User code
+ Add New Item
Output Data
Data
MC Data
Data and software details
Documentations
CADI ID
URL
Keyword
Comment
Internal Discussions
URL
Private and public documentation.
Presentations
URL
Publications
Journal Title
Journal Year
Journal Volume
Journal Issue
Journal Page
Identifiers
+ Add New Item
Automatic filling from existing DBs.

Open data portal

First official version in November 2014.



- Samples used for Masterclasses
- Simple event displays
- Histograms

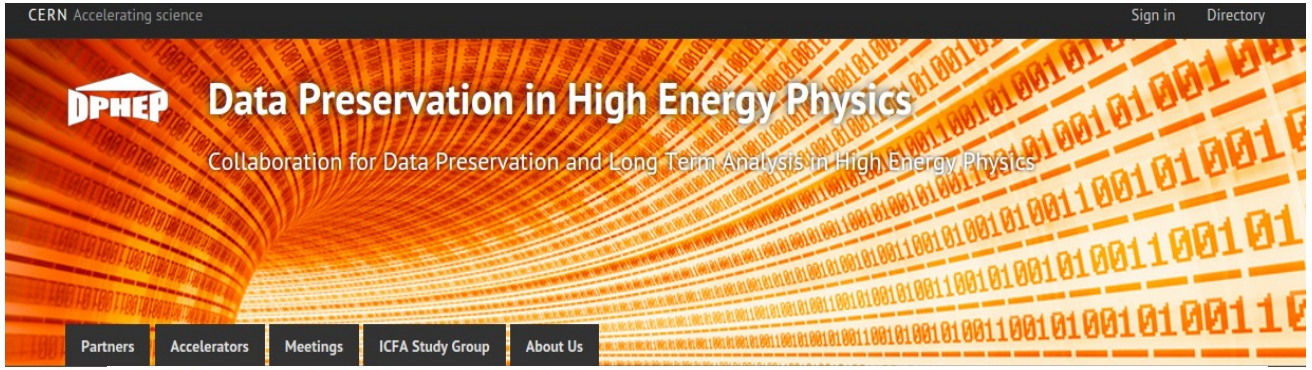
- Data used by the collaboration (analysis level samples, downloaded via xrootd)
- Experiment software (CernVM image)

Currently CMS only (2010: 28 TB, 2011: 130 TB) . *Other LHC experiments will join in the next future.*

DPHEP portal

http://hep-project-dphep-portal.web.cern.ch/

Portal to collect all DP project in past and current experiments.



FOLLOW THE LINKS BELOW TO FIND INFORMATION ON OUR PARTNER ORGANIZATIONS. EACH REPRESENT SOME EXPERIMENTS AND ACCELERATORS TO THE COLLABORATION FOR DATA PRESERVATION IN HIGH ENERGY PHYSICS.

BNL	Home	
CERN	Home	Data
CSC	Home	
DESY	Home	
Fermilab	Home	
IHEP	Home	
IN2P3	Home	
INFN	Home	
IPP	Home	
KEK	Home	

Search this site — Search

EXTERNAL RESOURCES

- Open Data Portal — A library of openly accesible physics data from CERN.
- HEPData — An open-access repository for scattering data from experimental particle physics.
- EUDAT — European infrastructure providing research data services.
- DPHEP Study Group — A common reflection on data persistency and long term analysis in High Energy Physics.

LHCb

Contact Information: Greybook

Public Page: http://lhcb-public.web.cern.ch/lhcb-public/

Internal Page: http://lhcb.web.cern.ch/lhcb/

The LHCb (standing for "Large Hadron Collider beauty") experiment is one of the particle p experiments collecting data at the Large Hadron Collider accelerator at CERN. LHCb is a s b-physics experiment, that is measuring the parameters of CP violation in the interactions o (heavy particles containing a bottom quark). Such studies can help to explain the Matter-An asymmetry of the Universe.

The detector is also able to perform measurements of production cross sections and electr in the forward region. Approximately 840 people from 60 scientific institutes, representing 1 form the collaboration which built and operate the detector.. The experiment is located at p LHC tunnel close to Ferney-Voltaire, France just over the border from Geneva. The (small) experiment shares the same cavern.

Bit Preservation:

Data and MC samples are stored on tape and on disk. Two copies of raw data on tape ; 1 c reconstructed data (FULL.DST, which contains also raw data) ;4 copies of stripped data (D last (N) reprocessing. Two copies for the N-1 reprocessing. One archive replica on tape.

Data:

For the long term future, LHCb plans to preserve only a legacy version of data and MC sa legacy data: 1.5 PB (raw), 4 PB FULL.DST, 1.5 stripped DST. Run 1 legacy MC : 0.8 PB D

Open data: LHCb plans to make 50% of analysis level data (DST) public after 5 years, 100 after it was taken. The data will be made public via the Open Data portal (http://opendata

Samples for educational purposes are already public for the International Masterclass Pro accessible also via the Open Data portal (For Education area).

Documentation:

Data: dedicated webpages for data and MC samples, with details about all processing step

Software : twiki pages with software tutorials, mailing-lists.

Documentation to access and analyse masterclasses samples is available on LHCb webpa OpenData portal.

Software:

BACKUP

DPHEP vision

The "vision" for DPHEP – first presented to ICFA in February 2013 – a consists of the following key points:

- By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – should be easily **findable** and fully **usable** by the **designated communities** with clear (Open) access policies and possibilities to annotate further
- Best practices, tools and services should be well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
- There should be a **DPHEP portal**, through which data / tools accessed
- **Clear targets & metrics** to measure the above should be agreed between **Funding Agencies, Service Providers** and the **Experiments (Collaborations).**

Open data policies @ LHC

Level 1 (published data)	All scientific results are public. Additional data associated with the results will also be made available (e.g. Histograms data); additional info archived in Inspire or HEPdata.
Level 2 (samples for educational purposes)	Data samples in simplified format for event displays and masterclass exercises.
Level 3 (reconstructed data)	CMS will release **50% of their total data 3 years after data taking** LHCb will release **50% 5 years after data is taken, 100% after 10 years.** Alice will release **10% after 5 years, 100% after 10 years.** Atlas will release their data **after a embargo period**.
Level 4 (raw data)	Due to the complexity of the raw data processing stage, the extensive computing resources required and enormous access to tape resources, direct access to raw data is not permitted to individuals within the collaboration. Raw data processing is performed centrally. Due to this, CMS/LHCb are currently **not planning to allow open access to raw data**