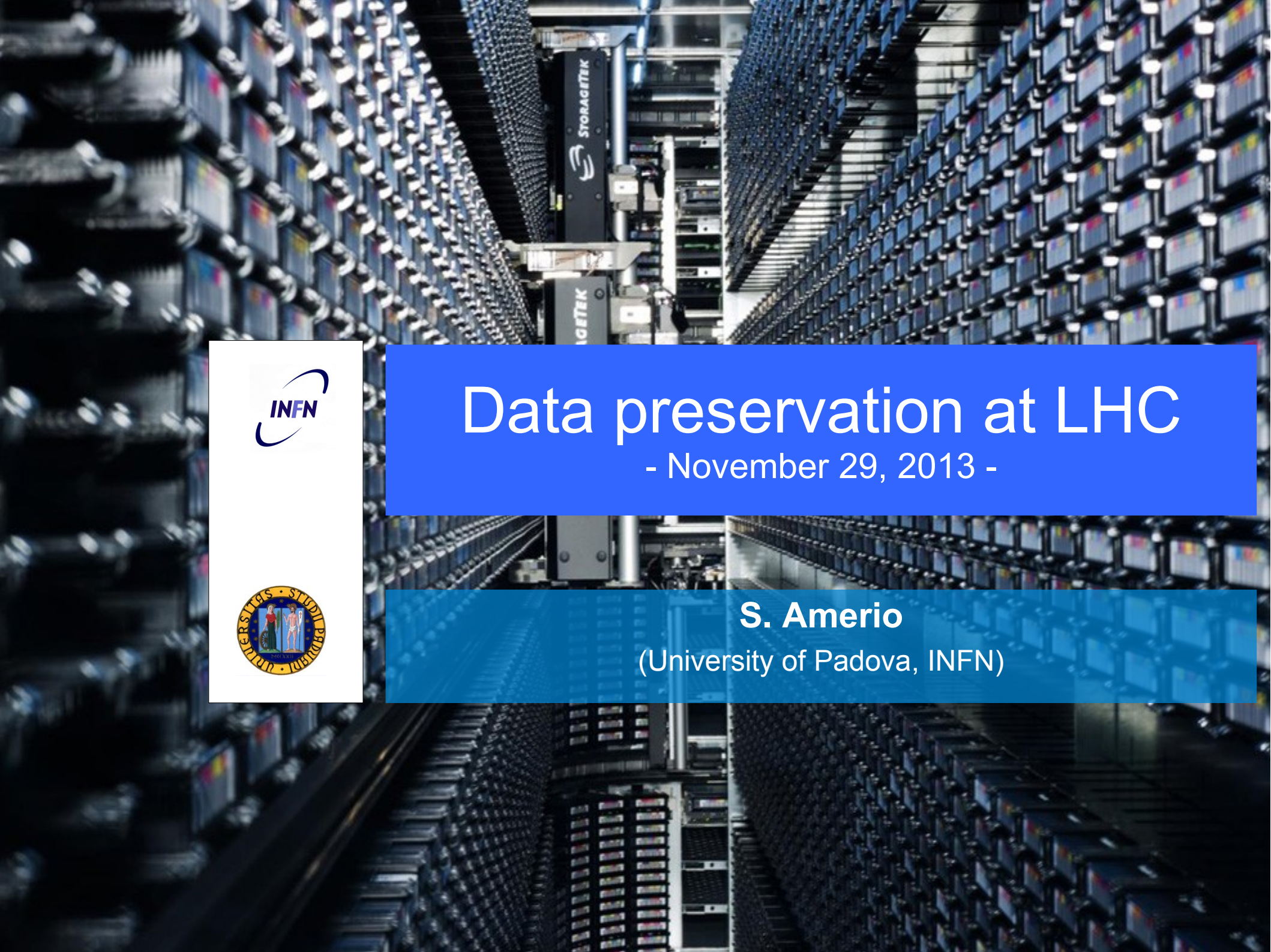


Data preservation at LHC

- November 29, 2013 -

S. Amerio

(University of Padova, INFN)

Introduction

Data preservation activities are becoming more relevant in HEP experiments.
Long term data preservation (and open access) matters for

- social reasons: our experiments are paid by the community, data should be preserved and made available to general public.
- scientific reasons:
  - old data can always be re-analyzed in search of new signals or to improve measurements.
  - scientific results should be reproducible.

All LHC experiments are now actively working on data preservation issues.
- Dedicated task forces
- Dedicated sessions/talks at collaboration meetings/workshops
- DP activities are considered as service tasks to the experiment

Common areas of work:

- Open access and outreach
- Bit and software preservation
- Validation systems
- Documentation/Analysis preservation

Open access

CMS open access policy official since March 2012, LHCb since March 2013
Atlas/Alice: drafts under discussion within their collaborations.

CMS/LHCb policy (based on the DPHEP levels of data preservation):

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 data 3 years after data taking** LHCb will release **50% 5 years after data is taken, 100% after 10 years.**
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**

Open access

CMS open access policy official since March 2012, LHCb since March 2013
Atlas/CMS: drafts under discussion within their collaborations.

CMS/LHCb policy (based on the DPHEP levels of data preservation):

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 data 3 years after data taking** LHCb will release **50% 5 years after data is taken, 100% after 10 years.**
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. Centrally. Due to this, CMS/LHCb are **open access to raw data**

It means we need to provide external users with
- Access to data
- Software to analyse data
- Documentation

Open access project in CMS

- Preparing for a public release of part of 2010 collision and simulated data in AOD format, appropriate for physics analysis, and accompanied by stable, open source software needed for a number of example analysis and suitable documentation.
- A virtual machine (VM) image has been prepared, in a format usable by the freely-available VirtualBox application.
- For the access to the data, the initial work flow is kept as close to the standard one as possible, which uses xroot. An xrootd server has been commissioned with anonymous read-only access, further limited by firewall to include only those sites involved in the testing phase.
- Preparing high-school level classroom applications chosen as a pilot use-case[6] - tools to process the data and documentation into an appropriate format, studies on:
  - a common definition of data contents for most HEP experiments
  - a common HEP ontology with Linked Data methods.

From Kati Lassili-Perini poster at CHEP2013

Open data project funded by Finnish government in collaboration with IT Center for Science. The project is part of a larger plan of Finnish Ministry of Education to put scientific data in use, and a pilot project for the platform of the open reserch material at CSC.
https://twiki.cern.ch/twiki/bin/view/HIPCMSExperiment/CMSOpenDataProject

Other open access & outreach projects

All LHC experiments have released samples of data in simplified format for educational purposes, to be used in event displays and mastrclass exercises.

Public Data

- Data Samples
- Analysis Tools & Formats
- CMS Masterclass
- Useful links

CMS Public Data

The CMS experiment at the LHC has released a portion of its data to the public for use in education and outreach. Explore this page to find out more about the data and how to analyse it yourself.

Try the online event display below.

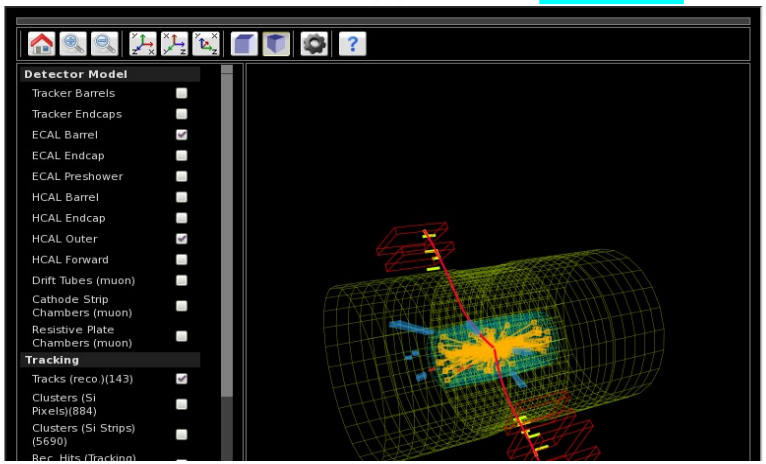
Use the Mouse to rotate.

Ctrl+Mouse or Ctrl + ▲ ▼ ◀ ▶ to pan x/y.

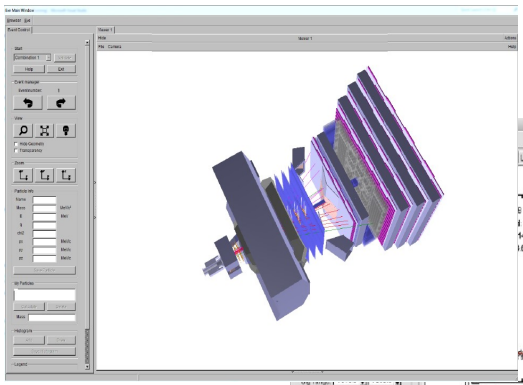
Shift+Mouse or Shift + ▲ ▼ to zoom.

Follow the links at left to access the full version of the display and thousands of ev

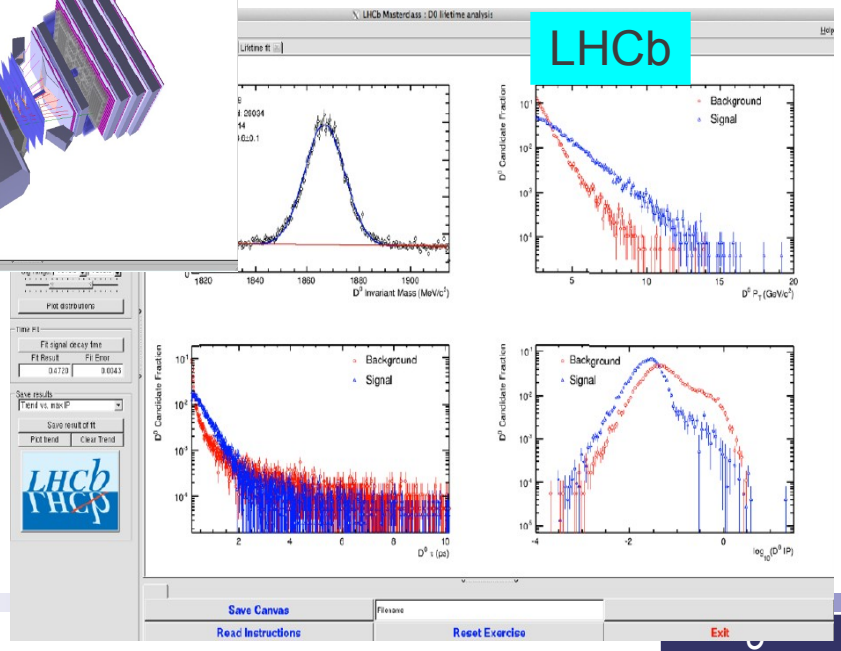
CMS



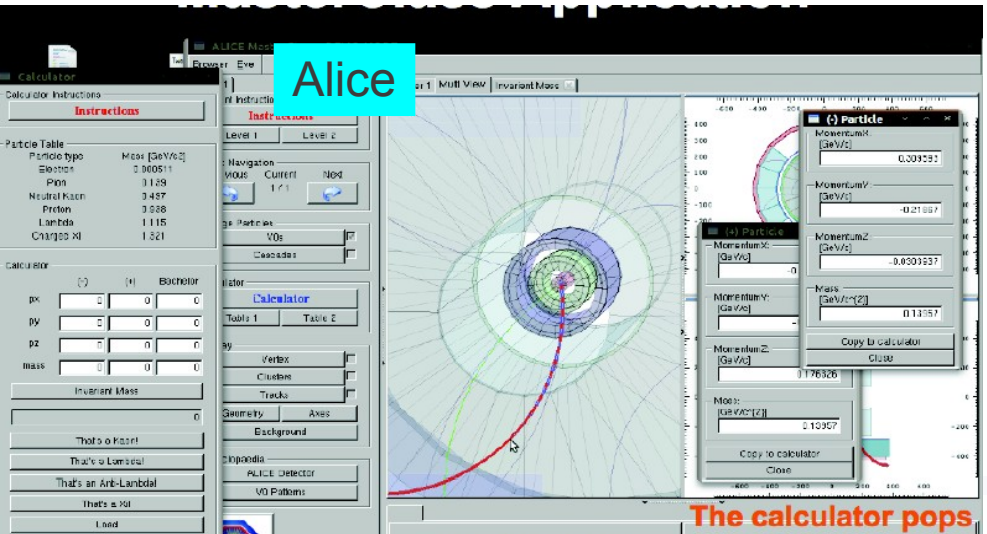
ATLAS



LHCb



Alice



DP @ LHC

Data and software preservation

CMS/LHCb
- All raw data preserved
- Define legacy datasets / software releases: all Run I data reprocessed into a legacy dataset with a single software version
- Use virtualization techniques to run legacy software
- Backup policies
- Data integrity checks

- Analysis level software preserved in dedicated repositories

Alice/Atlas developinig similar plans

Long term validation

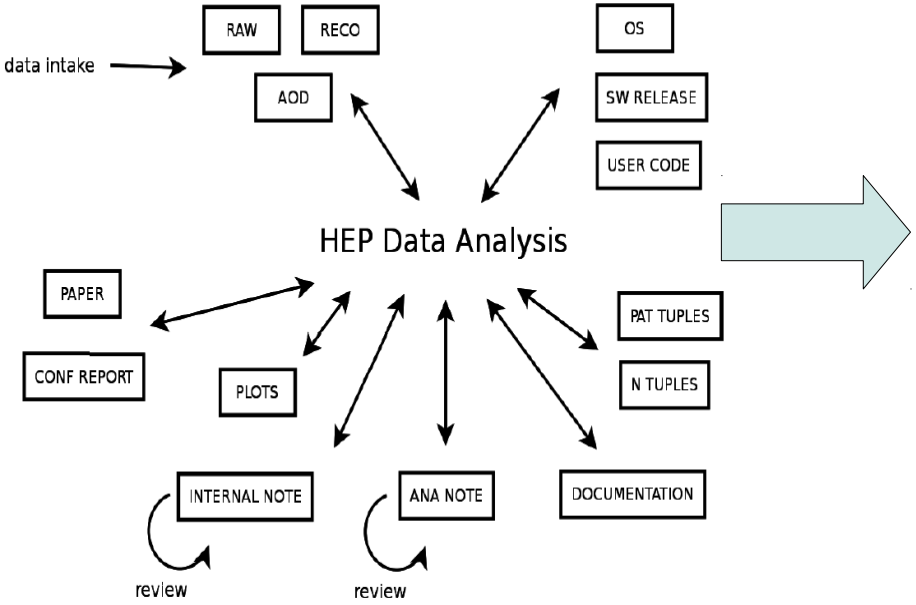
An efficient validation system is a fundamental ingredient of any long term data preservation plan.
Archived data and software need to be regularly checked against data loss and/or corruption, new operating systems and hardware.
LHC experiments are building long term validation systems *on top of their current validation tools.*

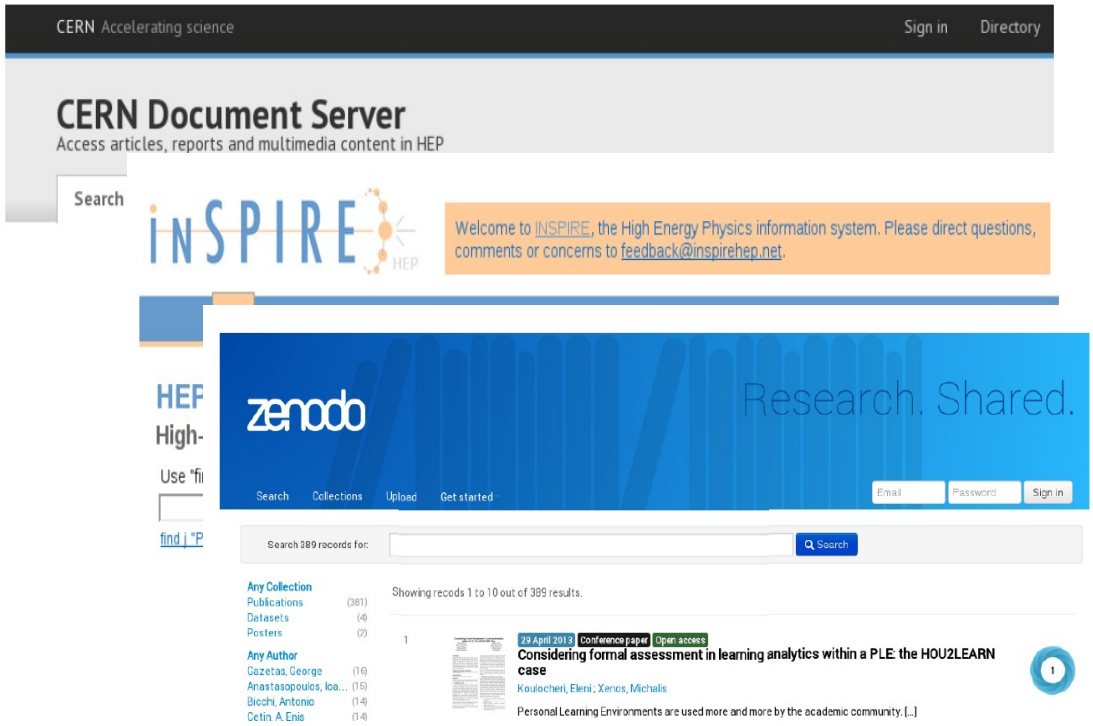
CMS: a set of reference plots being dened covering physics results: from individual objects (muons, electrons,...) to high level physics signatures involving basic analysis selections.

LHCb: for each step of data processing define long term future validation “analysis” in the current LHCb Performance & Regression tool

Analysis preservation

CMS/LHCb/Cern-IT
The Invenio team at CERN, with input from CMS and LHCb, will setup a prototype of a tool to make recording and documenting of the workfow and intermediate data easy.





Atlas
RECAST: framework developed to extend impact of existing analysis; it allows an existing analysis to be reinterpreted under an alternate model hypothesis --> complete information from original analysis needs to be preserved.

Conclusions

LHC experiments are actively working on data preservation.

They are working within DPHEP; collaboration is encouraged but not organized so far.
Same problems to face --> a lot of room for collaboration/common projects.
At the moment not discussed yet within WLCG.

The final goal should be the development of a common framework for long term preservation of HEP data.

- Backup -

Study group started in 2009
Blueprint released in May 2012 (available at dphep.org)
Priorities:

- *Secure data in all experiments*
- *Consolidate the on-going international collaboration*
- *Promote common multi-experiment projects and/within interdisciplinary cooperation*

DPHEP is now moving from a study group to an organization (collaboration)

- **Full time project manager** (J.Shiers) to coordinate the activities
- Definition of a **collaboration agreement** to be signed by laboratories and experiments → already signed by Cern.
- Regular meetings with representatives from all lab/experiments to discuss ongoing DP projects.

DPHEP vision:

By 2020 all archived data easily findable, fully usable by designated communities with clear (open) access policies.

Best practices, tools and services well run-in, fully documented and sustainable; built in common with other disciplines, based on standards.