HEPData:
the long-term data preservation facility in particle physics

Frank Krauss
IPPP Durham University

"Our Digital Future"

Multidisciplinary Perspectives on Long-Term Data Preservation and Access
Cambridge 15.6.2016

HEPDATA

IP3
www.ippp.dur.ac.uk

Long-term data preservation (in HEP): why is this an issue

- LHC and other particle physics experiments constitute a sizeable and important investment of money and time – their full exploitation is part and parcel to honouring the investment
- to maximise their scientific impact it is paramount to
  - store the data in a robust format that can be migrated, thereby enhancing their lifetime
  - keep them openly and freely available for everybody;
  - allow their re-interpretation;
  - allow training of a new generation of particle physicists with them
- this is our scientific legacy and we should be proud of it

Long-term data preservation in HEP: what do we want to preserve?

- for sake of argument assume a long gap – 50 years - in HEP experiments (no direct transfer of knowledge, experience and expertise between generations of PhD students).
- what do people need in order to built the "LHC 2100"?
  - real events – they must gain intuition of what is expecting them! these will essentially be raw data
  - high-level, analysed data in their context, for training, validation of new tools and procedures
- I will focus on the second kind of data and their preservation

HEP publications

HEP Scattering experiments going back to the 1950s

Each group of scientists will analyse particular signals by processing large numbers of collision.

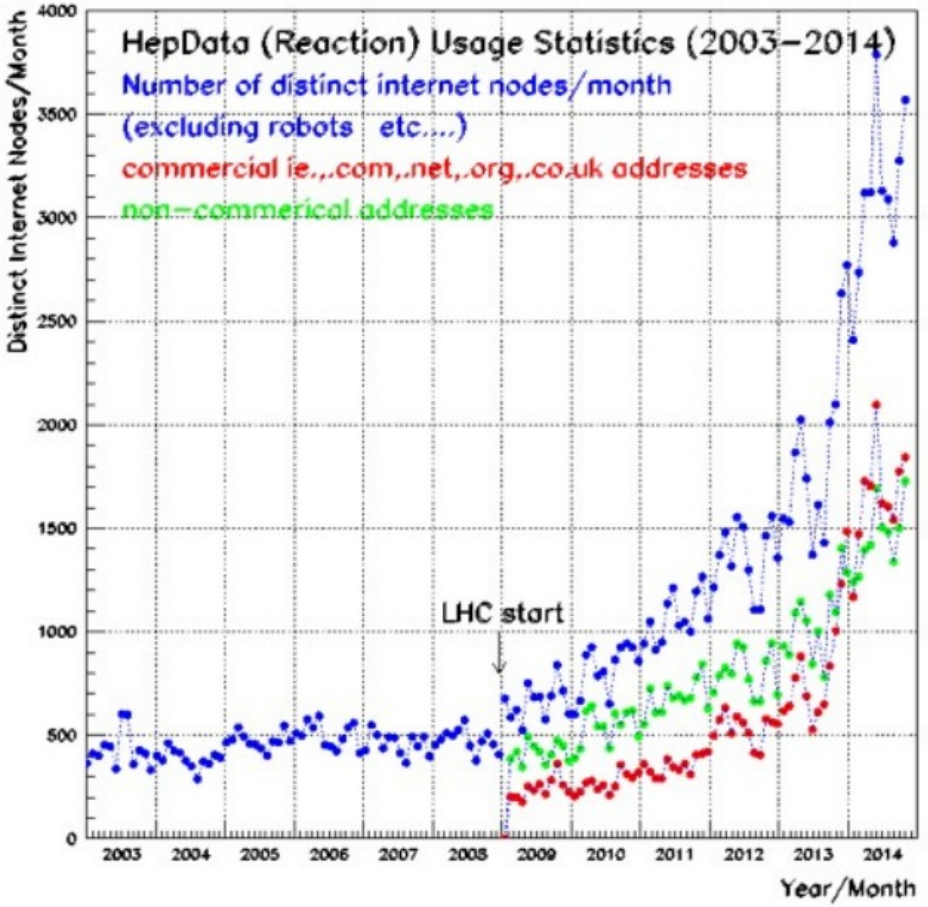
The resulting analysis will be published as a paper.

But where to keep the data accessible?

HEPData's role

- a unique, persistent & up-to-date database for results of experimental particle physics beyond the lifetime of the experiments
- located and developed by a small team (2 people) at IPPP Durham since nearly 40 years
- funded by STFC as part of the experimental programme (and part of their data strategy)
- hosts about 64000 datasets from more than 8000 papers, often supplemented with additional information
- all datasets are stored as numbers, in a modern database format (to allow detailed comparison with theory or similar)

HEPData usage



Details of stored data

- results from ~8000 papers

  (January 2015: ATLAS: 164/294, CMS: 83/284, ALICE: 74/87, LHCb: 23/49)

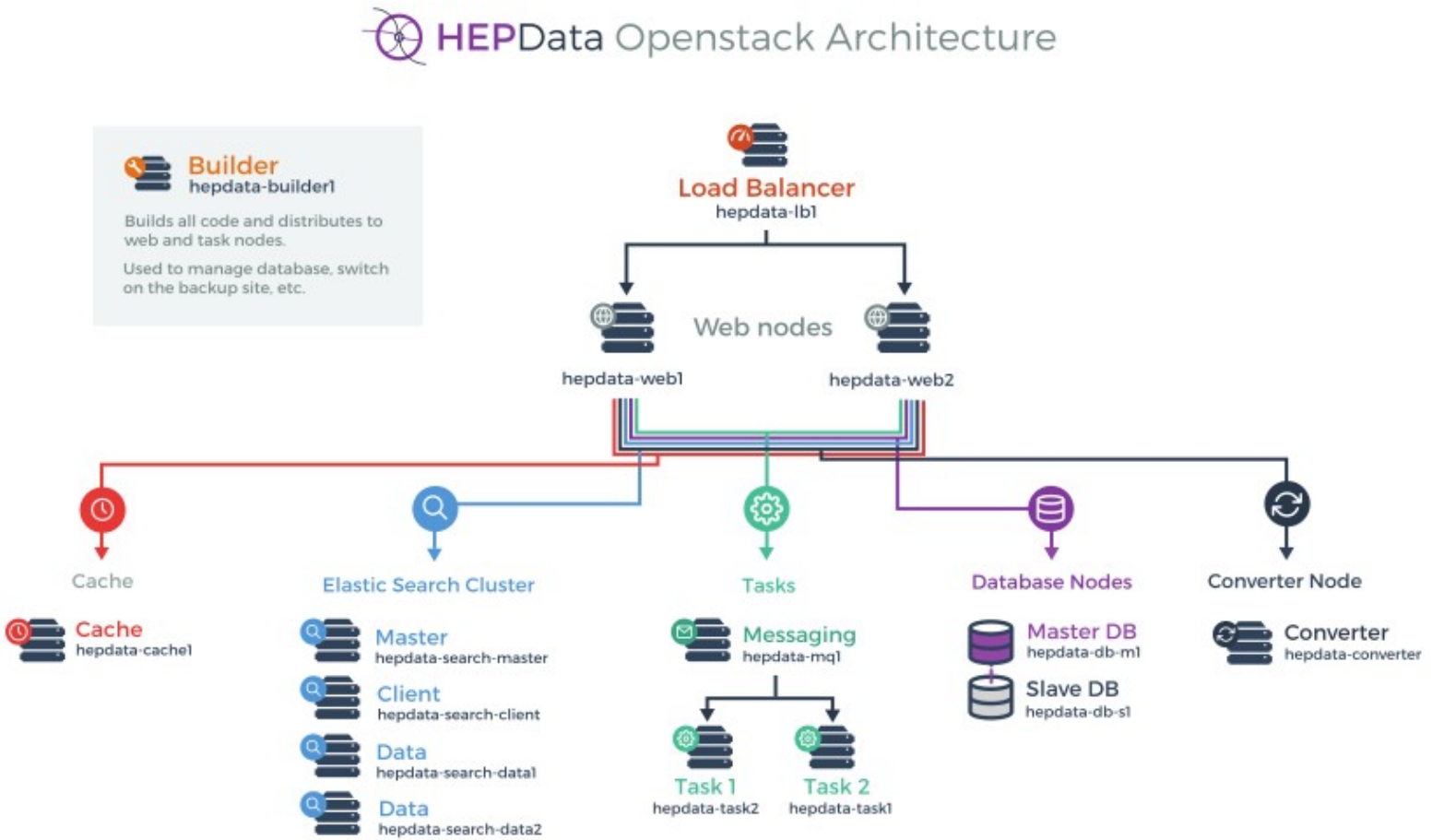
  mostly of Standard Model cross sections and other measurements of scattering experiments dating back over 40 years

- systematic error breakdown, correlation matrices, SLHA files etc.
- linked through web pages, different output formats, including plots
- used for many purposes: for example MC tuning, input for or facilitating new measurements
- in the past: manual upload of data by HEPData team, not sustainable in LHC era due to volume of publications → **self-upload mandatory**

HEPData 2.0: new technology

- **migration of HEPData to INSPIRE/Invenio**
  - INSPIRE is central database for HEP publications, people, experiments, jobs, … runs on Invenio database
  - similar mission, similar user basis, identical vision
  - freeing time for HEPData for further developments & better service to the community
  - freeing time for developing improved strategies of data publication, curation, preservation, and discovery
  - first step towards an integrated long-term strategy of data preservation, providing better contextual information

Machinery of HEPData in Invenio

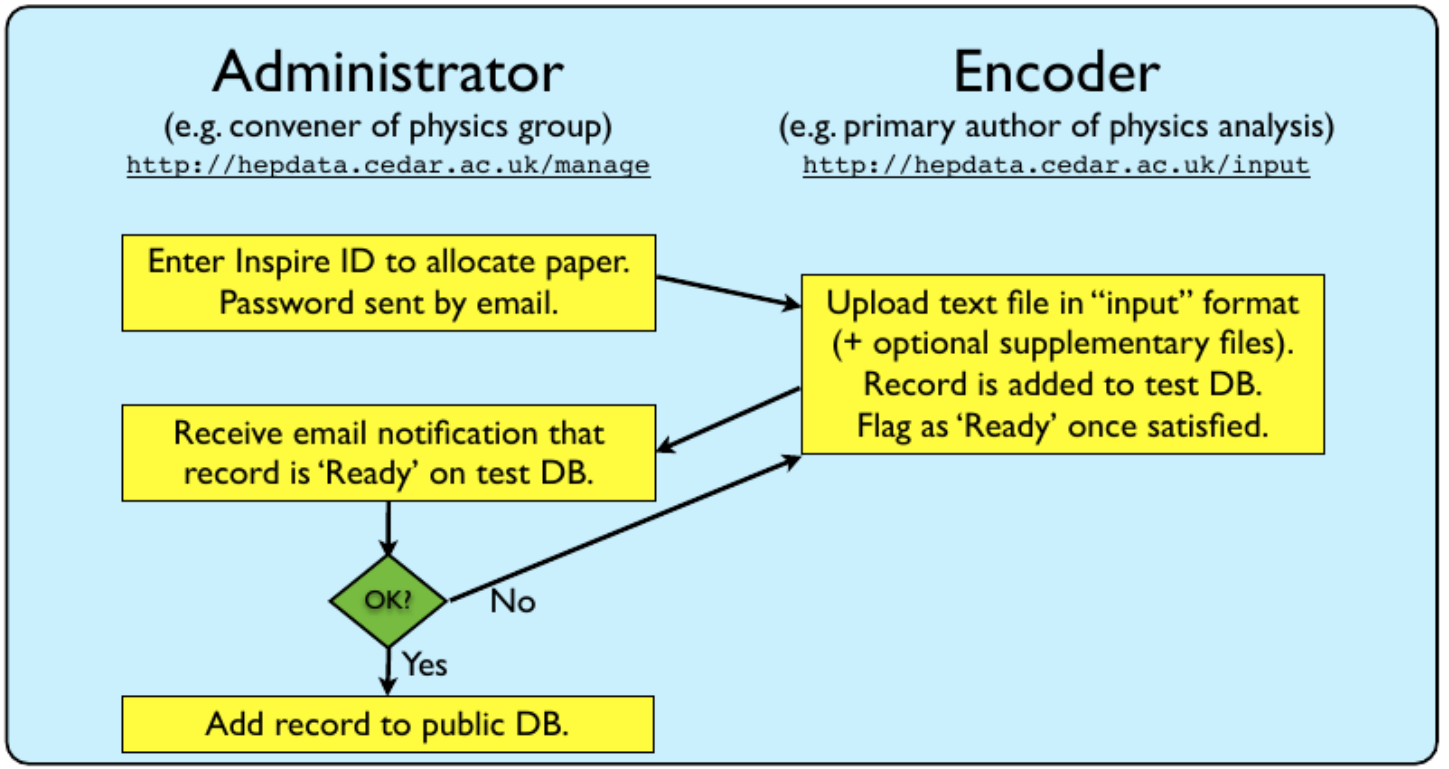


HEPData 2.0: new services

- re-implement and vastly improve search system
- better methods for inclusion of supporting material and cross listings
- direct upload of data in ROOT format
  (Advisory Board tasked us with this! trickier than we thought)
- extend the scope of HEPData
  - particle decays (b/c-hadrons, τ's) → under way
  - data for detector simulations
  - low-energy data, astrophysics, …
- remember: only 2 people really doing the job

Modus operandi for uploading data

- quality control of uploaded data:
  - correct → numbers in database agree with published data/plot
  - fast → ensure quick turnover to wider community and public
- keep intellectual responsibility and control with experiments



Modus operandi for uploading data: input format and look-and-feel

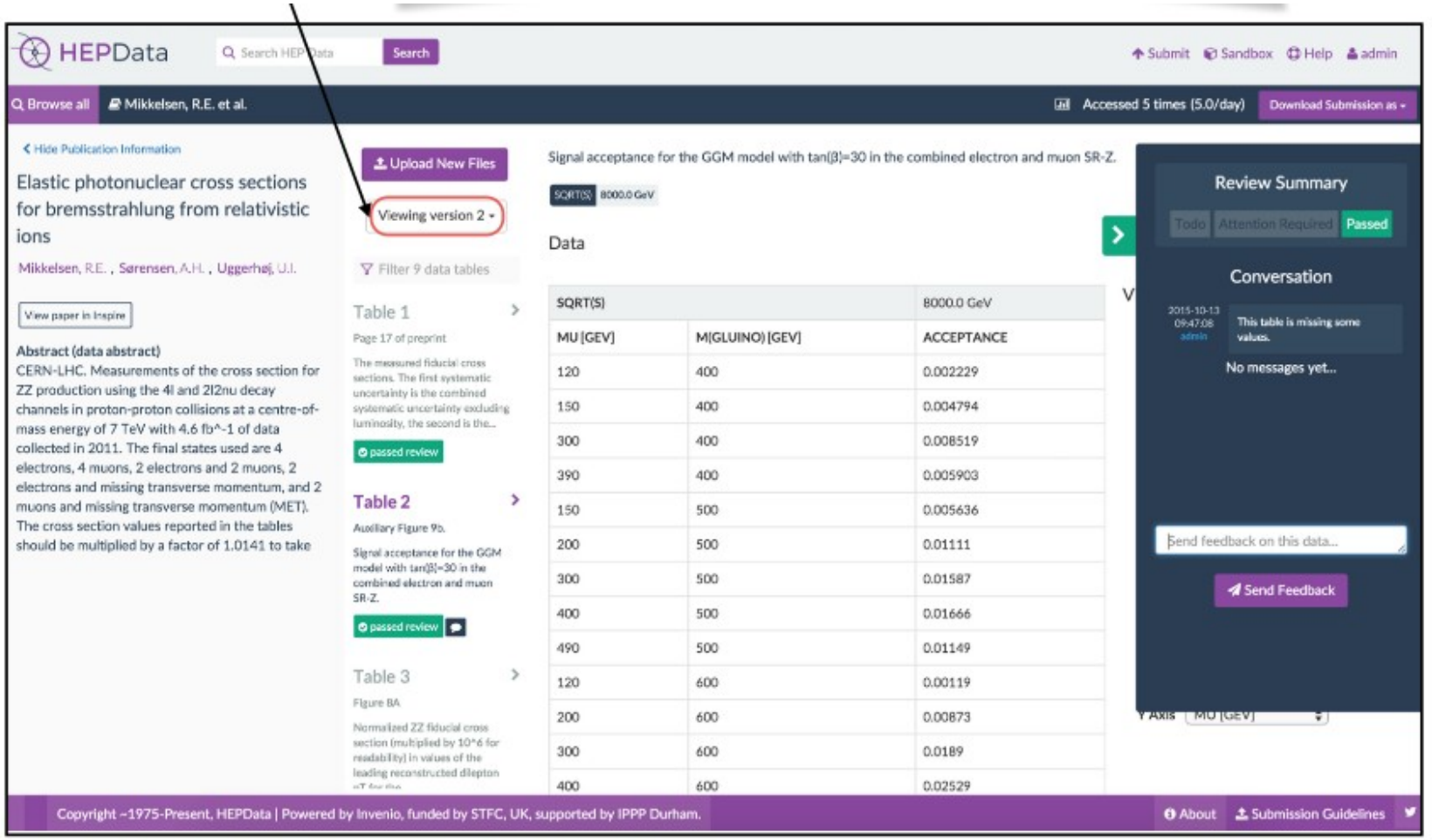
http://hepdata.cedar.ac.uk/view/ins1203852/d2/yaml

```
---
name: 'Table 2'
label: 'Data from Page 20 of preprint'
description: |
    The measured total cross sections.  The first systematic uncertainty is the
combined systematic uncertainty excluding luminosity, the second is the luminosity.
keywords:
  - {name: reactions, values: ['P P --> Z0 Z0 X']}
  - {name: observables, values: ['SIG']}
  - {name: cmenergies, values: [7000.0]}
additional resources:
independent_variables:
  - header: {name: 'SQRT(S)', units: 'GEV'}
    values:
      - {value: 7000}
dependent_variables:
  - header: {name: 'SIG(total)', units: 'FB'}
    qualifiers:
      - {name: 'RE', value: 'P P --> Z0 Z0 X'}
    values:
      - value: 6.7
        errors:
          - {symerror: 0.7, label: 'stat'}
          - {asymerror: {plus: 0.4, minus: -0.3}, label: 'sys'}
          - {symerror: 0.3, label: 'sys,lumi'}
```

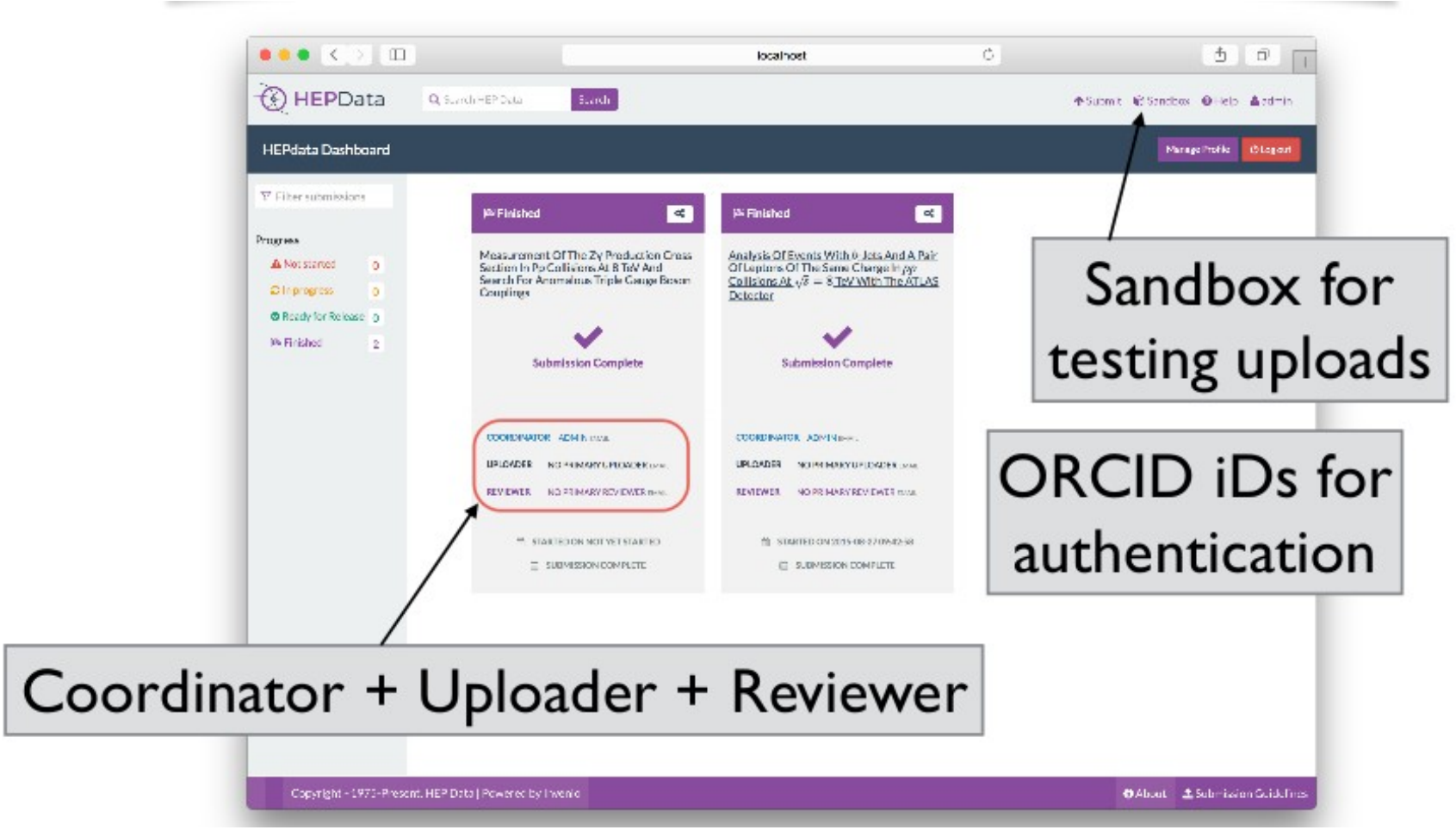
http://hepdata.net/record/60079

RE	P P --> Z0 Z0 X
SQRT(S) [GEV]	SIG(total) [FB]
7000	6.7 ±0.7 stat -0.3, 0.4 sys ±0.3 sys,lumi

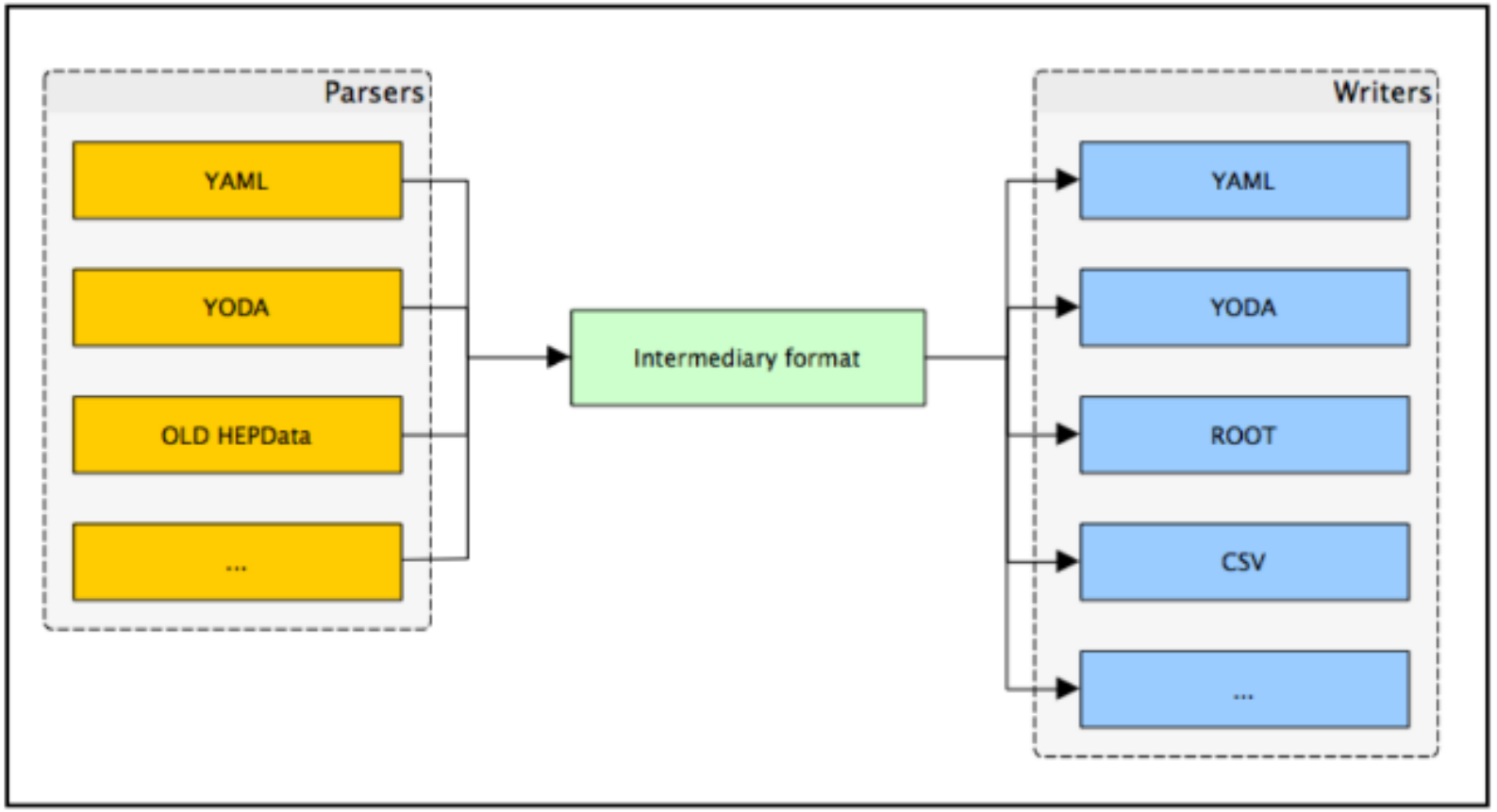
Modus operandi for uploading data: versioning and reviewing



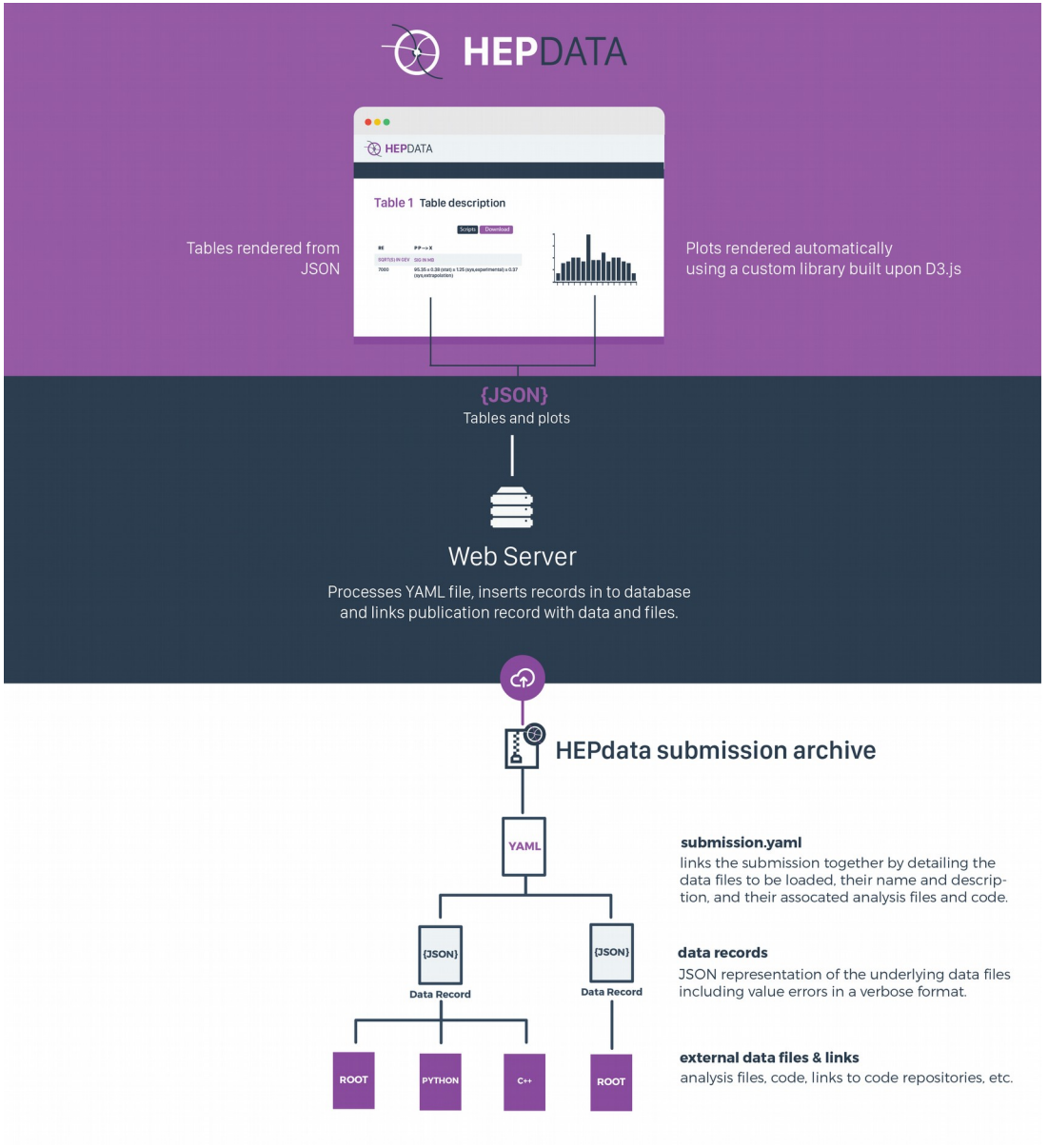
Modus operandi for uploading data: dashboard and steering the flow



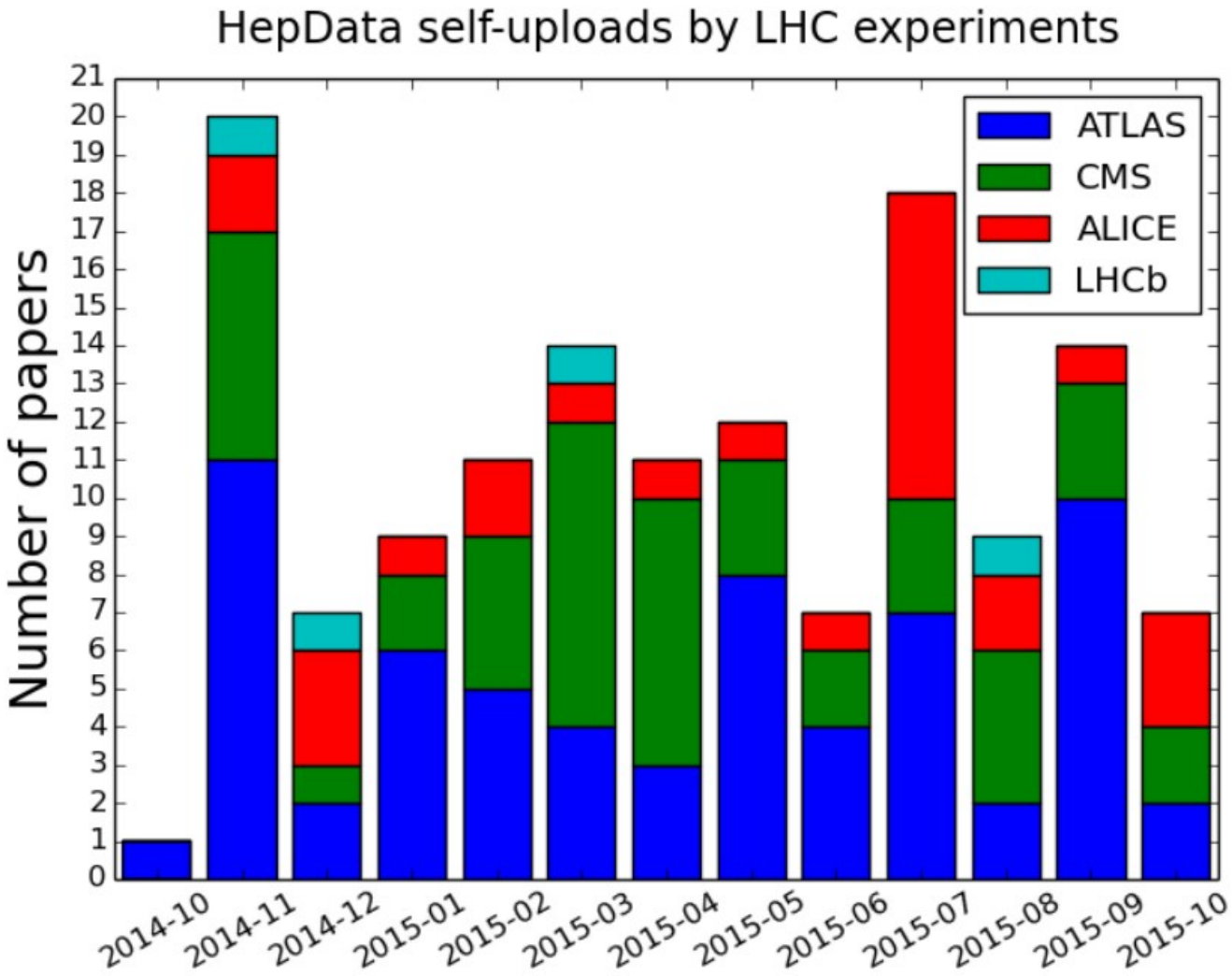
Modus operandi for uploading data: automatic format conversion



Organisation of uploads



Volume of self-uploads from LHC in first year of new system



Effect of new self-upload system

- role of HEPData manager changed from “uploader” to “curator”
- main tasks now:
  - data curation: improving searchability, developing methods/strategy for embedding and contextualising data
  - improving user interface
  - quality control of data by spot checks and fine-tuning upload procedure for experiments
  - widening the scope of database: used to be scattering only, started to include decay data and data relevant for detector construction

The first τ decay paper in HEPData

The Durham HepData Project — Durham University

REACTION DATABASE • DATA REVIEWS • PDF PLOTTER | ABOUT HEPDATA • SUBMITTING DATA

Reaction Database Full Record Display

View short record or as: input, plain text, AIDA, PyROOT, YODA, ROOT, mpl, ScaVis or MarcXML

BUSKULIC 1997 — A study of τ decays involving η and ω mesons

Experiment: **CERN-LEP-ALEPH (ALEPH)**
Published in **ZP C74,263 (1997)** (DOI:10.1007/s002880050387)
Preprinted as **CERN-PPE-96-103**
Preprinted as **FSU-SCRI-97-50**
Record in: **INSPIRE**

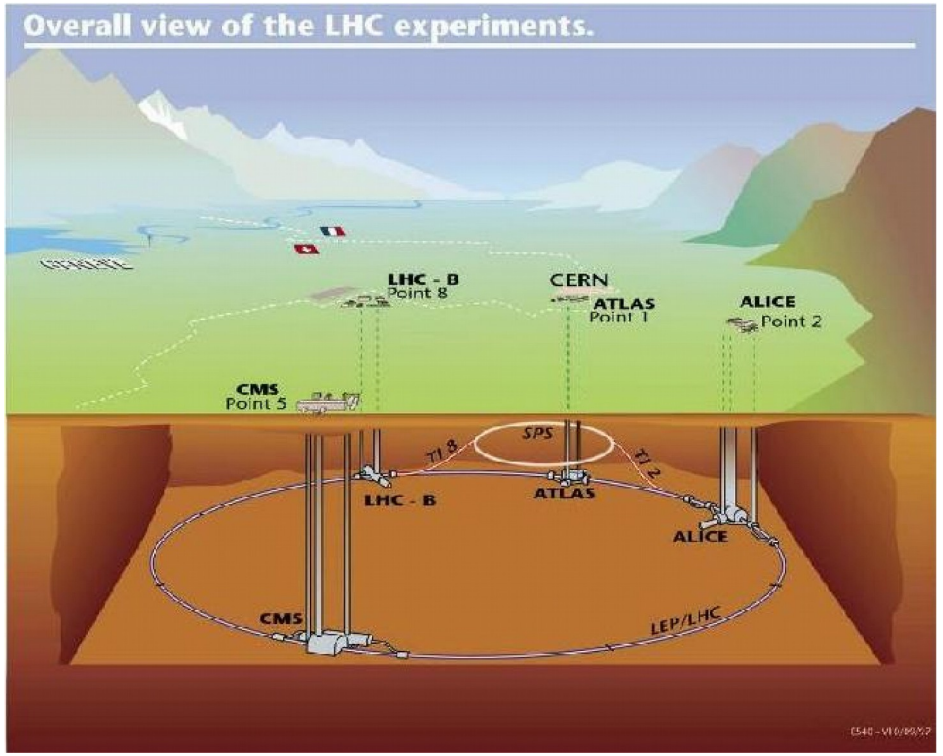
CERN-LEP. The 132 pb^{-1} of data collected by ALEPH from 1991 to 1994 have been used to analyze η and ω production in τ decays. The following branching fractions have been measured:

$$B(\tau^- \to \nu_\tau \omega h^-) = (1.91 \pm 0.07 \pm 0.06) \times 10^{-2},$$
$$B(\tau^- \to \nu_\tau \omega h^- \pi^0) = (4.3 \pm 0.6 \pm 0.5) \times 10^{-3},$$
$$B(\tau^- \to \nu_\tau \eta K^-) = (2.9^{+1.3}_{-1.2} \pm 0.7) \times 10^{-4},$$
$$B(\tau^- \to \nu_\tau \eta h^- \pi^0) = (1.8 \pm 0.4 \pm 0.2) \times 10^{-3}$$

and the 95% C.L. limit $B(\tau^- \to \nu_\tau \eta \pi^-) < 6.2 \times 10^{-4}$ has been obtained. The $\omega\pi^-$ and $\eta\pi^-\pi^0$ rates and dynamics are found in agreement with the predictions made from e^+e^- annihilation data with the help of isospin invariance (CVC).

These numbers have been read from the plots in the paper.

Aside: data taking at LHC



- typical for HEP: independent experiments (detectors) at the same accelerator
- run by large collaborations (e.g. ATLAS: ~3500 people from ~150 institutions in ~35 countries)
- all collaboration members usually signing authors

Aside: data taking at LHC

- disclaimer: this is a bird's eye view from a theorist
- data at the LHC is taken by detectors with a huge frequency and volume (hundreds of Gbit/s) – too much to be all stored
- there are various levels of selection criteria – triggers
- the stored data are affected by the detector (various detecting elements, electronics digitising the inputs, etc.)
- this limits the ability for direct comparison with theory and introduces an extra layer of uncertainty/error to be dealt with (through procedures known as “unfolding”, “detector corrections”)
- for measurements and their publication: must deal with such effects on different levels of sophistication

A vision for long-term data preservation: reproducible data

- where possible, define physical objects & correct for detector effects
- must clearly & unambiguously document object definitions & analysis
  → often code is better, clearer, and easier to migrate than papers
  → in contrast many publications are incomplete, assumptions are implicit, and in general language is subject to interpretation
- analysis code often only exists inside huge experiment-specific software (which is typically not openly available)
  → need experiment-independent analysis framework and
  → modular analysis to become part of the data, need validation
- we already have a framework, must provide infrastructure for analysis library and validation suite

HEPData vision of future procedures

- submit paper to arXiv
- put data + supplementary data/code/etc. into HEPData
- "public" area in RIVET etc. to add relevant code
- central repository for code and its validation
  - validation for example by adding MC run card & MC data and direct comparison
  - could be done in HEPForge (another library) or similar
- **upload all simultaneously: paper, data, code**

→ publish measured data + interpretation