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

Frank Krauss
IPPP Durham University

HEPDATA

IP3
www.ippp.dur.ac.uk

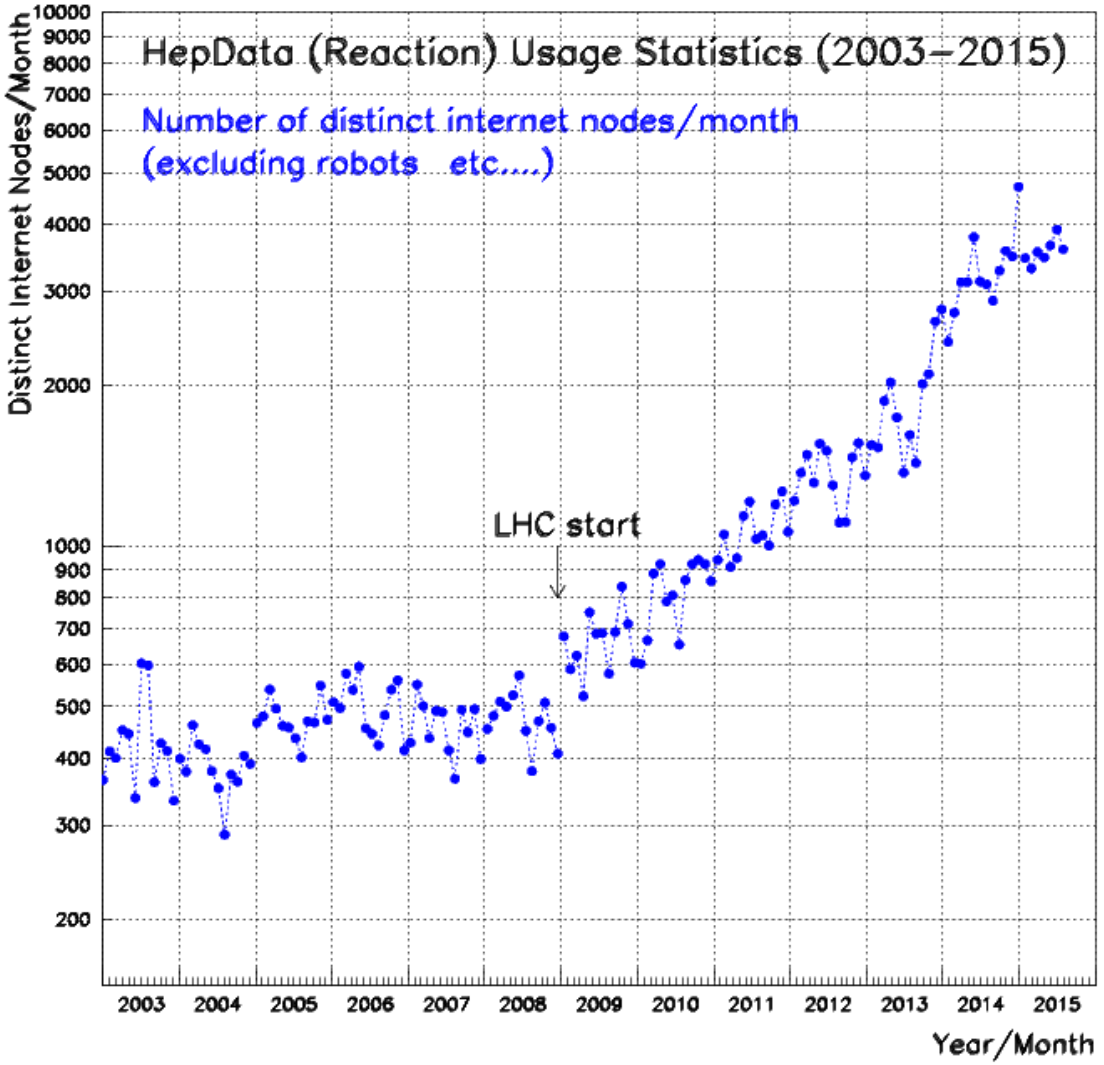
Long-term data preservation (in HEP): why is this an issue

- 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

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
- in past 1.5 years: move to Invenio/Inspire with massive support by CERN Inspire team
- funded by STFC as part of the experimental programme (and part of their data strategy)
- hosts about 65000 data tables 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

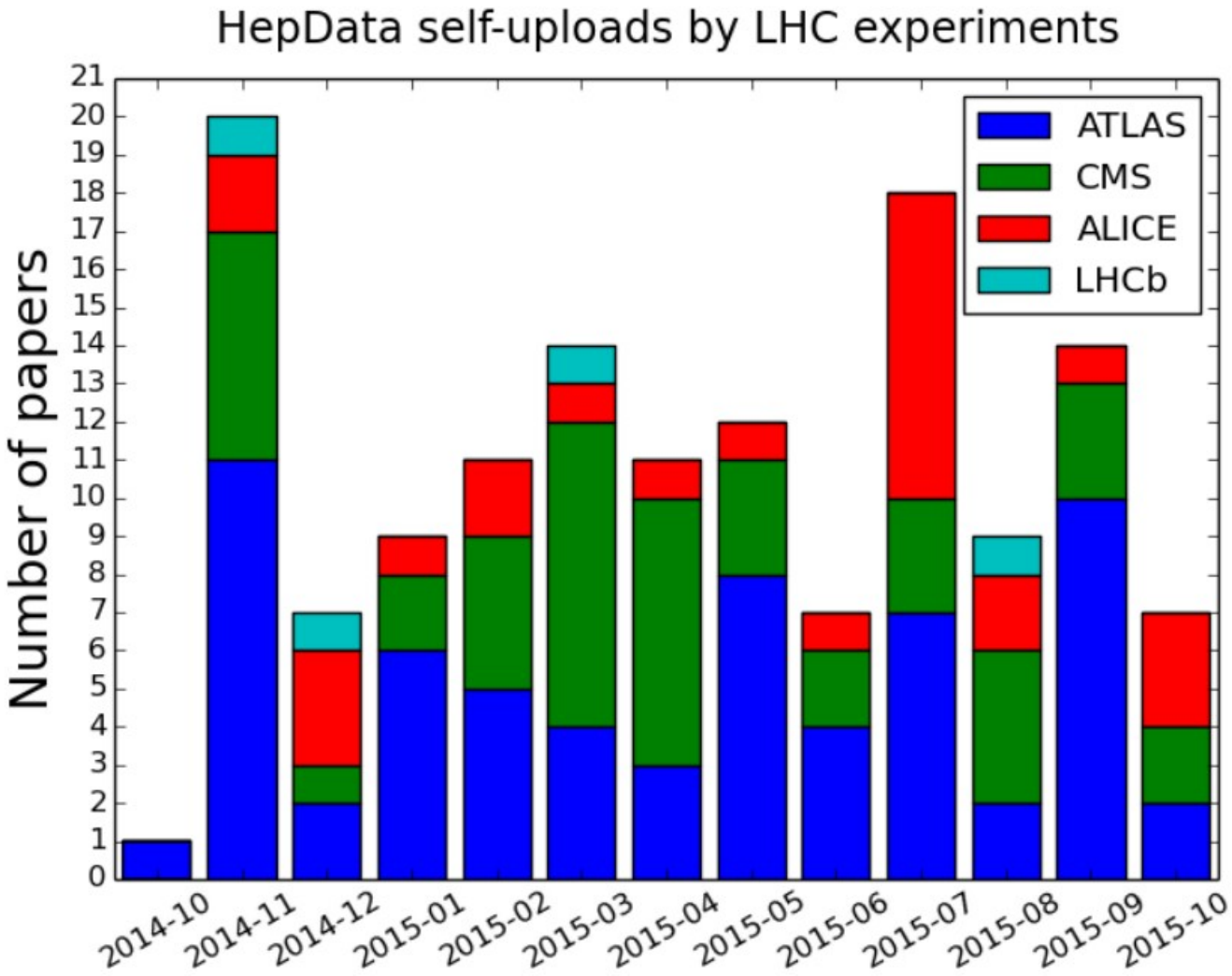
  (April 2016: ATLAS: 249, CMS: 162, ALICE: 116, LHCb: 37)

  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 preparation upload of data by HEPData team, not sustainable in LHC era due to volume of publications

  → **self-upload mandatory**

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

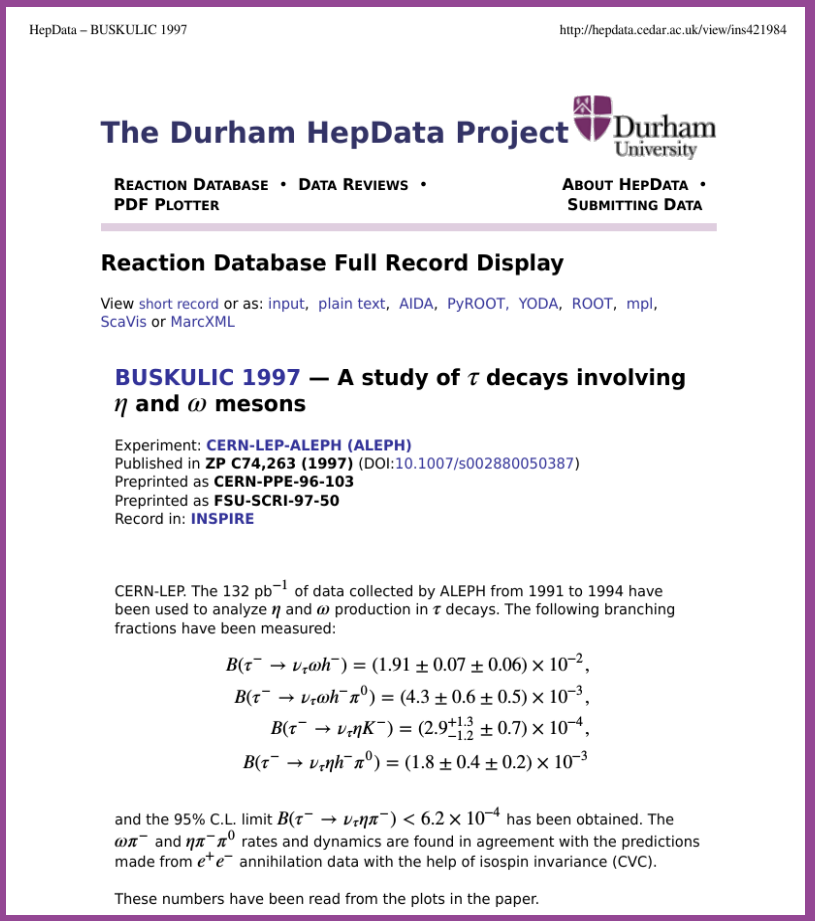
HEPData 2.0: new technology

- **migration of HEPData to INSPIRE/Invenio**
  - INSPIRE is central database for HEP publications, people, experiments, jobs, … runs on Invenio database
  - similar mission, similar user base, 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

HEPData 2.0: new services

- re-implement and vastly improve search system
- better methods for inclusion of supporting material
- extend the scope of HEPData
  - particle decays (b/c-hadrons, τ's) → under way
  - data for detector simulations → t.b.d. in next few months
  - low-energy data, astrophysics, ...

First tau paper in HEPData



HepData – BUSKULIC 1997 http://hepdata.cedar.ac.uk/view/ins421984

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.

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
- alternative: “public” area in RIVET etc. to add relevant code
- central repository for code and its validation
- **upload all simultaneously: paper, data, code**

→ publish measured data + interpretation

Governance of HEPData:

- Day-to-day business driven by PI and HEPData manager (G Watt)
  - 100% funded by UK – we need to report to STFC
- International Advisory Board:
  - annual one-day meetings: reporting of progress in HEPData, feedback from experiments and wider user community
  - discussing and advising on long-term strategy and short-term goals
- Members of IAB
  - 4 LHC experiments: Bill Murray (ATLAS), Henning Flaecher (CMS), Ulrik Egede (LHCB), Enrico Scomparin (ALICE)
  - Wider user community: Andy Buckley (Glasgow), Kyle Cranmer (NYU), Matthew Wing (UCL)
  - Inspire: Sunje Dallmeier-Tiessen, Salvatore Mele