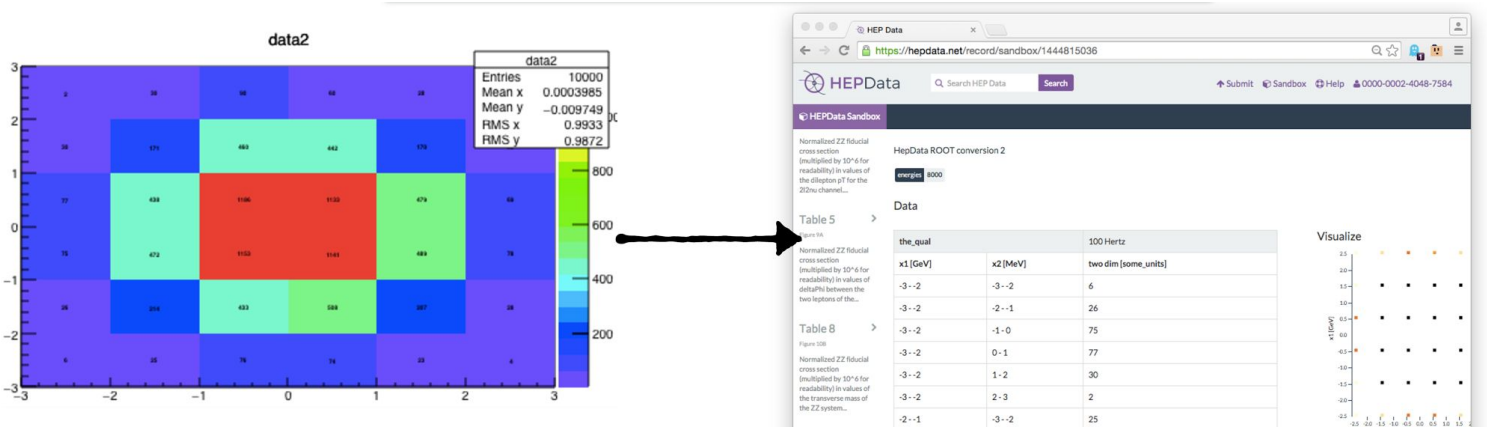
Additional Data Structures for HEPData

Lukas Heinrich 2016/06/15
Reinterpretation Workshop

HEPData Status Quo

HEPData can fulfill two roles:

1) **Destination**
   Serve as a long-term repository for quantitative data that is part of HEP experiments' research output: e.g. cross-sections, observed / expected event counts, observable distributions, limit contours, etc.
   Requirement:
   data stored in format independent of experiments. Experiments can either directly provided in HEPData format by or digested via in-house conversion code. More powerful conversion in the works: e.g. ROOT to lower barrier to entry.



HEPData Status Quo

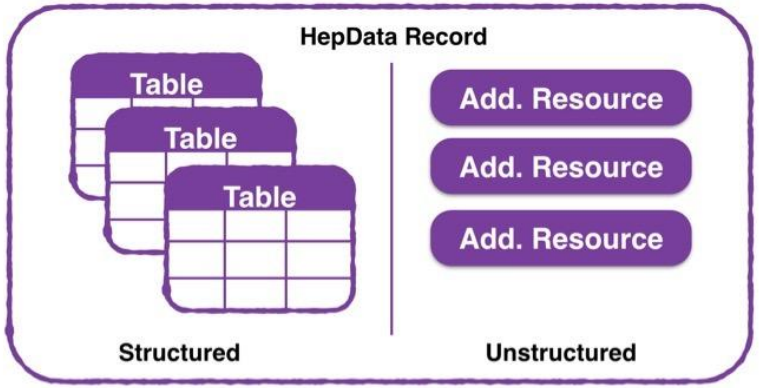
HEPData can fulfill two roles:

2) **Interface between hep-ex and hep-ph.**
   Data stored in HEPData can be starting point for new projects:
   a) HEPData Data Query Engine: query records in data-driven way. Useful for e.g. getting overview of existing searches ("show me all records with neutralino limit contours"). Promising, would benefit from more dataset uniformity (model parameter names etc)
   b) Store data that's suitable for **re-use** for e.g. reinterpretation/recasting projects: e.g. storing full likelihoods functions L(θ|data)

Current HEPData records tries to capture everything in a **single** data structure: the table, i.e. a multivariate function of indep. variables due to historical focus on tables in publications. Additional data can be added as "additional resource", but

- no format enforced
- wide variation between records

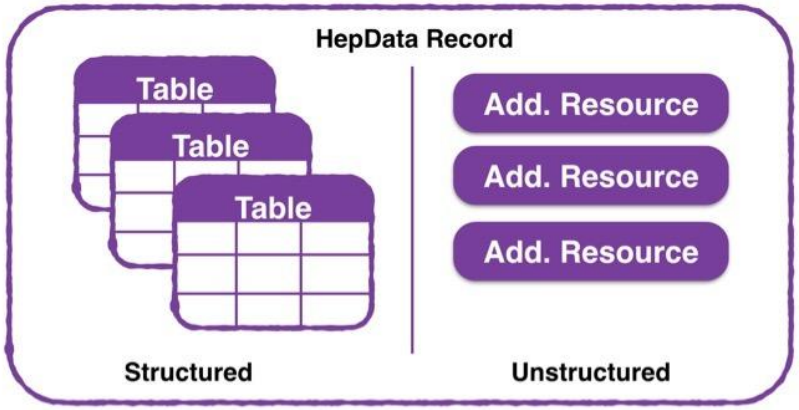
makes it difficult to use effectively.

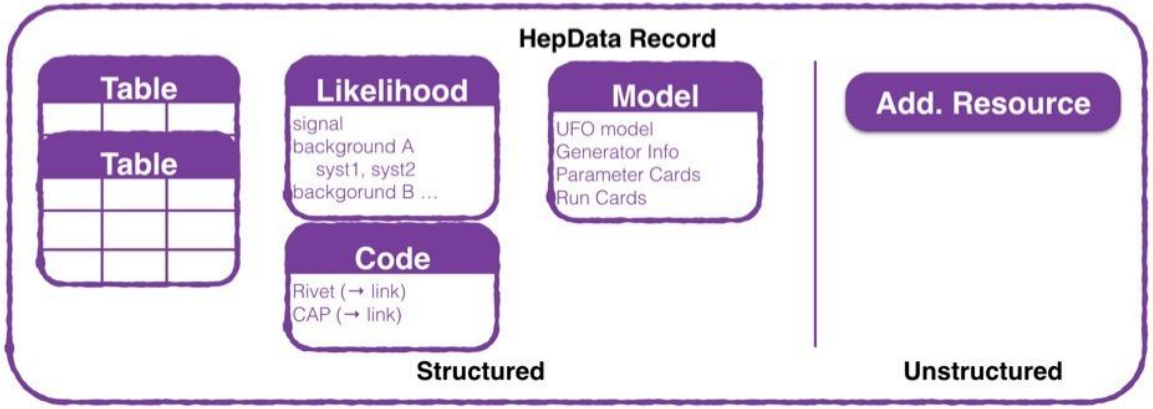


Beyond the Table

Moving beyond the Table

Not everything is naturally represented as a table. Have discussed to extend HEPData with a set of rich native data structures like likelihoods, full statistical models, uniform model information (e.g. UFO), references to related codes.

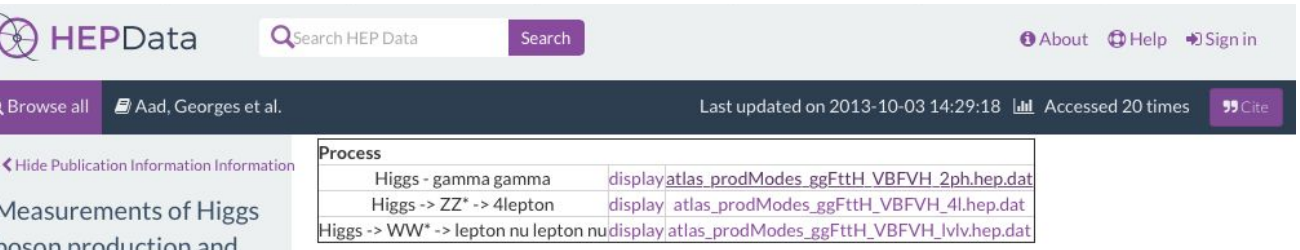


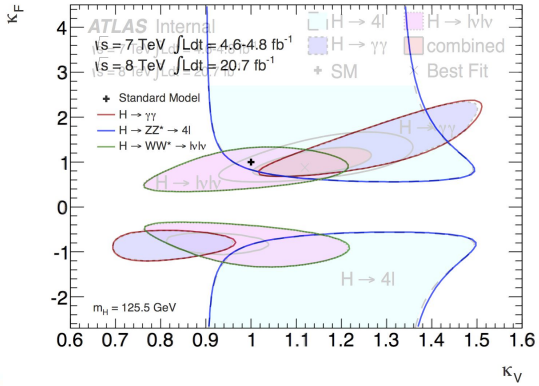


Data Structure Examples.

Some Examples:

1) **Likelihoods**
   publishing likelihood functions L(θ|data) allows continuously probing parameter points of the original model w.r.t. data
   - already published and cited as data products (see example), but not stored in structured way in HepData
   - it could already fit current format as multivariate function, but "Table" semantics don't really fit anymore.





NEW YC

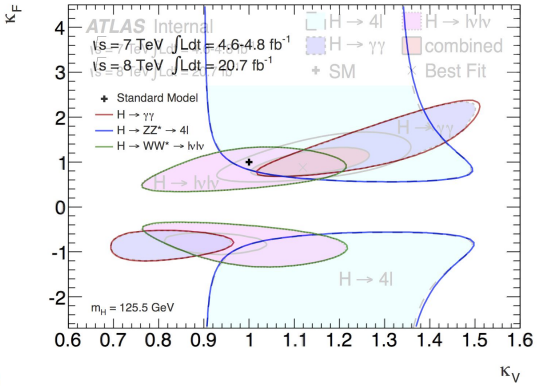
Data Structure Examples.

Some Examples:

1) **Likelihoods**

A more native description of likelihoods would simplify use of them for

- combinations of multiple analyses looking at same model
- re-interpretations, when new model can be expressed as function of old model parameters (e.g. coupling rescaling)
- Use-case example: export to Mathematica function



HEPData Search HEP Data Search About Help Sign in

Browse all | Aad, Georges et al. | Last updated on 2013-10-03 14:29:18 | Accessed 20 times | Cite

Hide Publication Information Information

Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

Process		
Higgs - gamma gamma	display	atlas_prodModes_ggFttH_VBFVH_2ph.hep.dat
Higgs -> ZZ* -> 4lepton	display	atlas_prodModes_ggFttH_VBFVH_4l.hep.dat
Higgs -> WW* -> lepton nu lepton nu	display	atlas_prodModes_ggFttH_VBFVH_lvlv.hep.dat

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP

Information | Citations (7) | Files

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC - ATLAS Collaboration

NEW YC

Data Structure Examples.

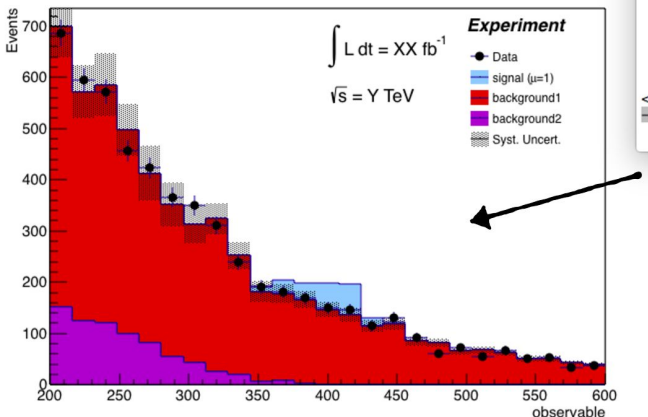
Some Examples:

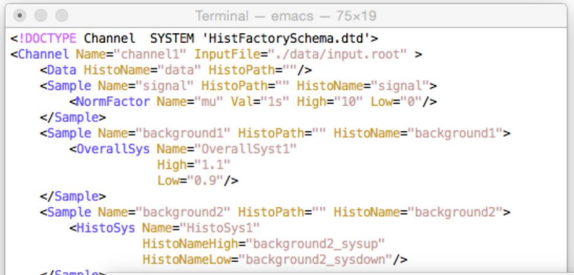
2) **Rich plot description via HistFactory**

- HistFactory is a tool widely used by experiments to built plots/p.d.f for binned distributions (i.e. histos), including full description of systematics. Uses 1↔1 mapping between pdf template and XML description + template histograms (in ROOT format)

$$\mathcal{P}(n_c, x_e, a_p \,|\, \phi_p, \alpha_p, \gamma_b) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c|\nu_c) \prod_{e=1}^{n_c} f_c(x_e|\boldsymbol{\alpha}) \right] \cdot G(L_0|\lambda, \Delta_L) \cdot \prod_{p \in \mathbb{S}+\boldsymbol{\Gamma}} f_p(a_p|\alpha_p)$$

- HepData can either use XML directly or convert to YAML (prototype exists) w/ Templates imported via HEPData ROOT converter.





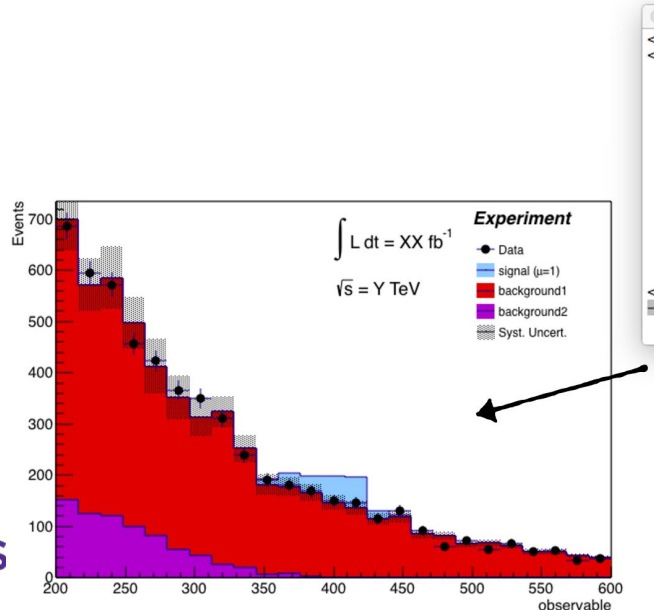
NEW YORK UNIVERS

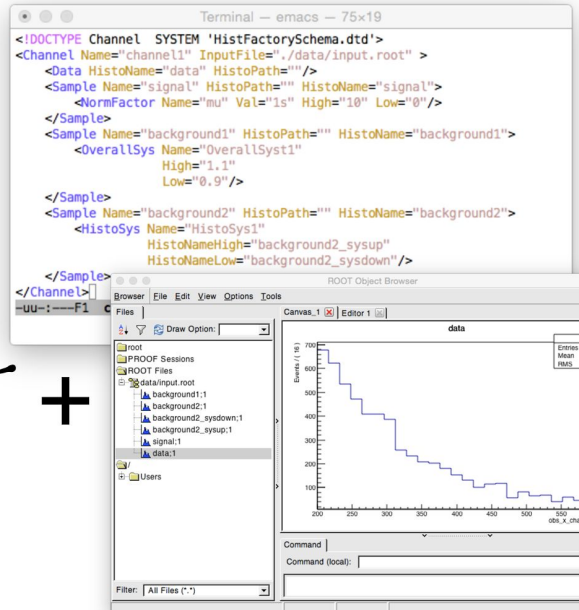
Data Structure Examples.

Some Examples:

2) **Rich plot description via HistFactory**

- Will give structured access to full set of systematic variation and background templates
- Tools and services like RECAST could provide new signal histogram





Data Structure Examples.

Some Examples:

3) **Uniform Model Descriptions**

- Information on models studied in particular publication hard to retrieve systematically. Mostly by pub title, sometimes SLHA files of grid points attached, but non-uniform/inaccessible to machines (limits e.g. HEPData Data Query Engine), adds to heterogeneity of HD dataset. Example: non-trivial to find all MSSM searches in HD
- Already nice partial solution by pheno community: a) SLHA files for SUSY, b) generic UFO models python modules that encode masses with generalized parameter cards.
- Would be very useful if HD had native notion of models, e.g. let experiments upload UFO(s) used in publications and assign identifier.

Data Structure Examples.

Some Examples:

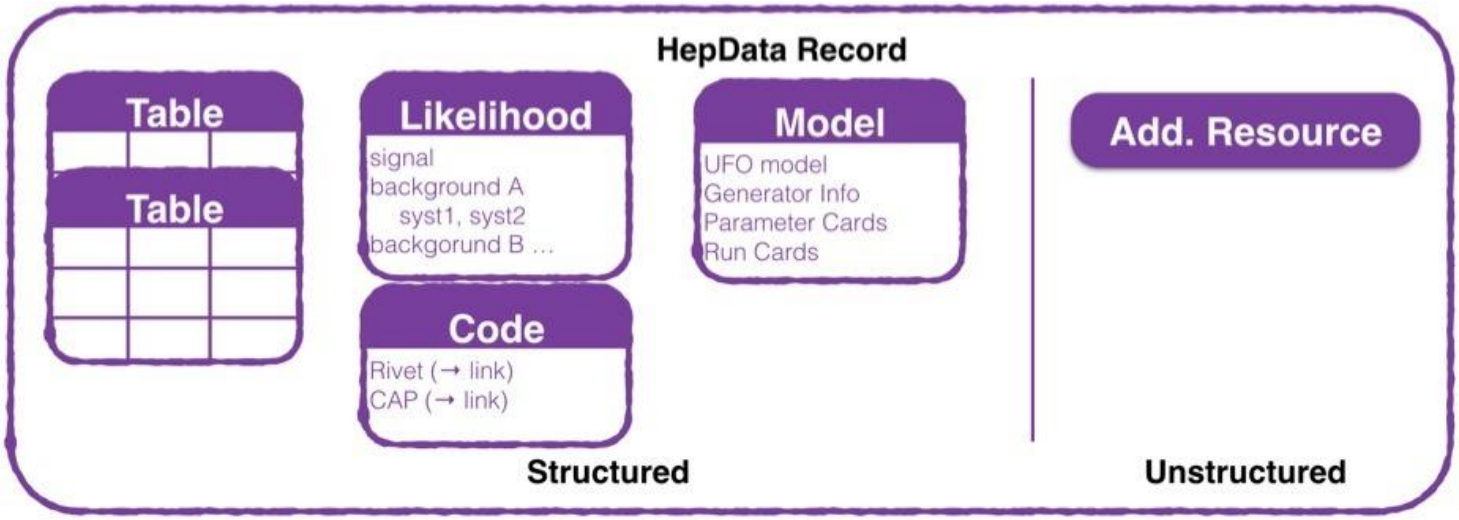
4) **Code/Catalogues cross-references**
   - publications can develop mini eco-system around themselves:
     i) Reimplementations via e.g. Rivet, CheckMATE, ATOM etc
     ii) Further information (e.g. analysis meta-data & code) in other catalogues like CERN Analysis Preservation Portal (soon!)
   - already captured in HD for Rivet in "add. resources", but could be extended to more sources, structured better / made more prominent (currently at same level as any link in add. resource of record)

Summary / Outlook

Launch of new HEPData an opportunity to revisit assumptions.

Good chance to expand utility of HEPData by expanding to more native data types beyond a simple table, both for higher fidelity archival of quantitative data of publications and its use for e.g. reinterpretation

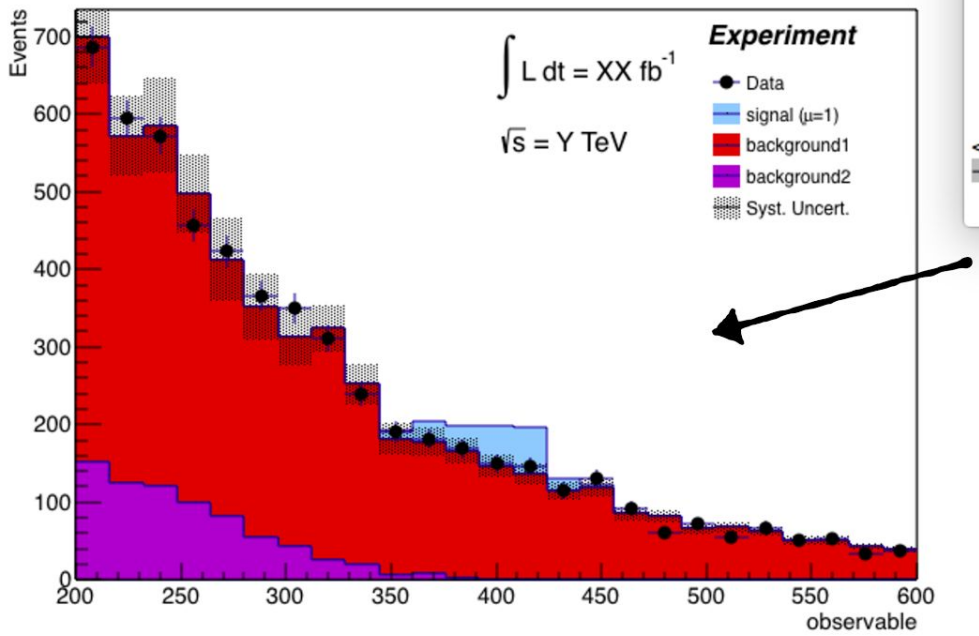
Doesn't need to come all-at-once, but new kinds of data types can be developed/added incrementally

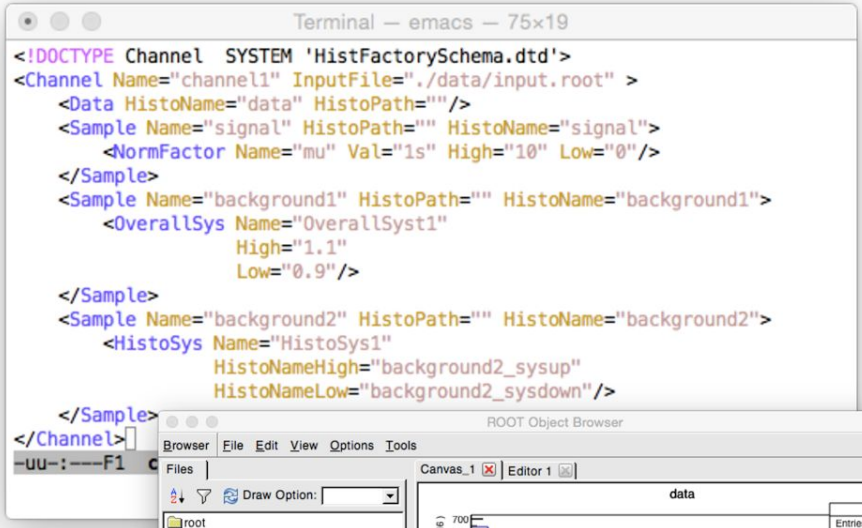


NEW YORK UNIVERSITY

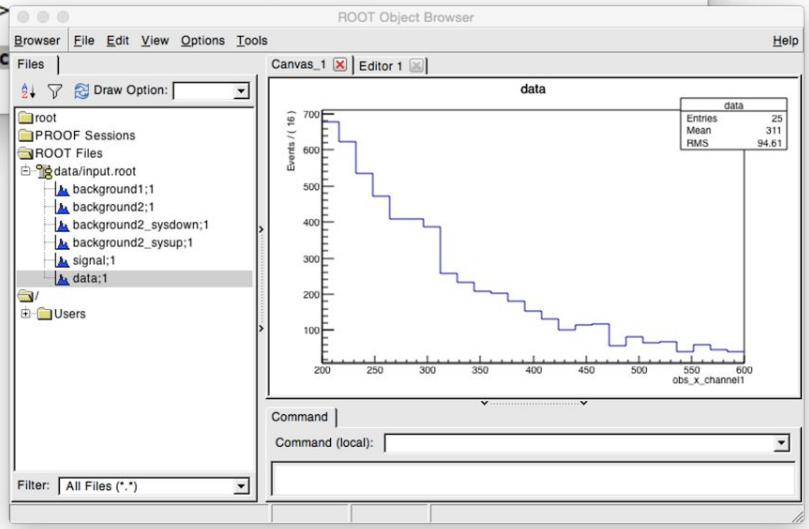
Backup

- HistFactory: model description framework
  - widely used in experiments, e.g. Higgs discovery, ~all SUSY analyses in ATLAS use HistFactory.
  - detailed description of distributions: resolved by component (signal, backgrounds) including systematic variations and correlations in well-formed schema.
  - format: (XML + ROOT)
  - underlying structure for many publication plots

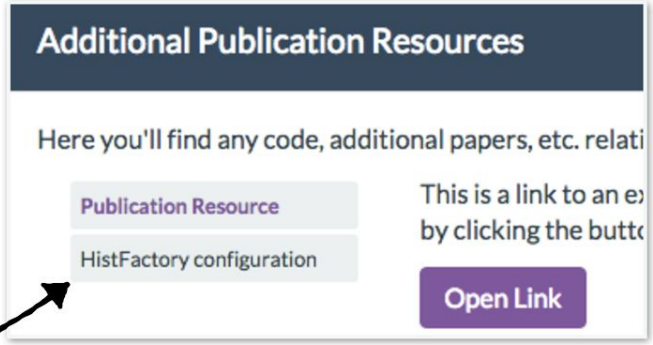




+



- HepData import
  - full HistFactory configuration (XML+ROOT) as “additional resource”
    - complete, machine-readable, exportable
  - developed set of tools to extract HepData tables from HistFactory configuration
    - result, uncertainties, correlations directly from measurement instead of manual compilation by experimenters.
  - increased coherence between publication and HepData record.



HistFactory