Integrating Scholarly Publications and Research Data - Preparing for Open Science, a Case Study from High-Energy Physics with Special Emphasis on (Meta)data Models

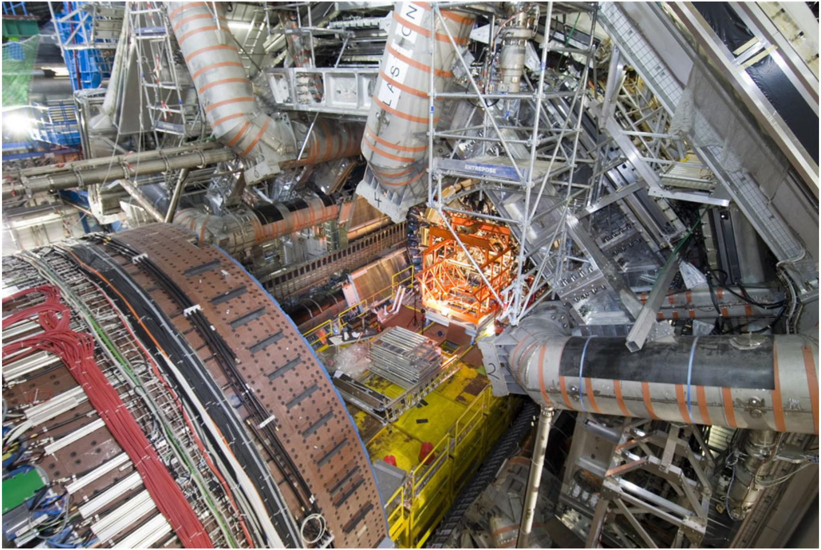
Piotr Praczyk[1,2] Javier Nogueras-Iso[2] Sunje Dallmeier-Tiessen[1] Mike Whalley[3]

[1]CERN, Geneva, Switzerland
[2]Computer Science and Systems Engineering Dept.,Universidad de Zaragoza, Spain
[3]IPPP, Durham University, UK

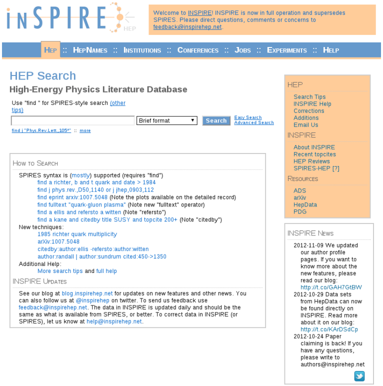
November 28, 2012

1. Introduction
2. Architecture of the integration
3. Benefits derieved from the integration
4. Conclusions and future work

High Energy Physics



INSPIRE



Welcome to INSPIRE! INSPIRE is now in full operation and supersedes SPIRES. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP Search

High-Energy Physics Literature Database

Use "find " for SPIRES-style search (other tips)

HOW TO SEARCH

SPIRES syntax is (mostly) supported (requires "find")

find a richter, b and t quark and date > 1984
find j phys.rev.,D50,1140 or j jhep,0903,112
find eprint arxiv:1007.5048 (Note the plots available on the detailed record)
find fulltext "quark-gluon plasma" (Note new "fulltext" operator)
find a ellis and refersto a witten (Note "refersto")
find a kane and citedby title SUSY and topcite 200+ (Note "citedby")

New techniques:

1985 richter quark multiplicity
arXiv:1007.5048
citedby:author:ellis -refersto:author:witten
author:randall | author:sundrum cited:450->1350

Additional Help:

More search tips and full help

INSPIRE UPDATES

See our blog at blog.inspirehep.net for updates on new features and other news. You can also follow us at @inspirehep on twitter. To send us feedback use feedback@inspirehep.net. The data in INSPIRE is updated daily and should be the same as what is available from SPIRES, or better. To correct data in INSPIRE (or SPIRES), let us know at help@inspirehep.net.

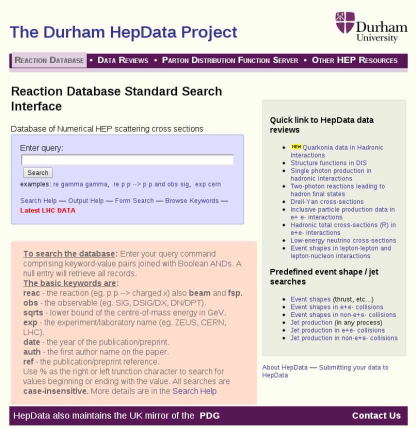
INSPIRE NEWS

2012-11-09 We updated our author profile pages. If you want to know more about the new features, please read our blog: http://t.co/GAH7GtBW
2012-10-29 Data sets from HepData can now be found directly on INSPIRE. Read more about it on our blog: http://t.co/KArDSdCp
2012-10-24 Paper claiming is back! If you have any questions, please write to authors@inspirehep.net

- The digital library of High Energy Physics (HEP) - Successor of SPIRES
- Over 10^6 records
- Harvesting from ArXiv.org, manual curation
- Added services: Reference counting, Author disambiguation, figures, ...

HepData

- Data behind around 7000 publications - over 50'000 datasets. (Data behind tables, plots and additional datasets)
- Manual curation
- Additional data submitted by authors of the publications
- Developed and maintained at the Durham University



The Durham HepData Project

Durham University
Reaction Database • Data Reviews • Parton Distribution Function Server • Other HEP Resources

Reaction Database Standard Search Interface

Database of Numerical HEP scattering cross sections

Enter query:

examples: re gamma gamma, re p p --> p p and obs sig, exp cern

Search Help — Output Help — Form Search — Browse Keywords — Latest LHC DATA

To search the database: Enter your query command comprising keyword-value pairs joined with Boolean ANDs. A null entry will retrieve all records.
The basic keywords are:
reac - the reaction (eg. p p --> charged x) also **beam** and **fsp.**
obs - the observable (eg. SIG, DSIG/DX, DN/DPT).
sqrts - lower bound of the centre-of-mass energy in GeV.
exp - the experiment/laboratory name (eg. ZEUS, CERN, LHC).
date - the year of the publication/preprint.
auth - the first author name on the paper.
ref - the publication/preprint reference.
Use % as the right or left trunction character to search for values beginning or ending with the value. All searches are **case-insensitive.** More details are in the Search Help

Quick link to HepData data reviews

- NEW Quarkonia data in Hadronic Interactions
- Structure functions in DIS
- Single photon production in hadronic interactions
- Two-photon reactions leading to hadron final states
- Drell-Yan cross-sections
- Inclusive particle production data in e+ e- interactions
- Hadronic total cross-sections (R) in e+e- interactions
- Low-energy neutrino cross-sections
- Event shapes in lepton-lepton and lepton-nucleon interactions

Predefined event shape / jet searches

- Event shapes (thrust, etc…)
- Event shapes in e+e- collisions
- Event shapes in non-e+e- collisions
- Jet production (in any process)
- Jet production in e+e- collisions
- Jet production in non-e+e- collisions

About HepData — Submitting your data to HepData

HepData also maintains the UK mirror of the **PDG** **Contact Us**

Fragment of a sample publication page

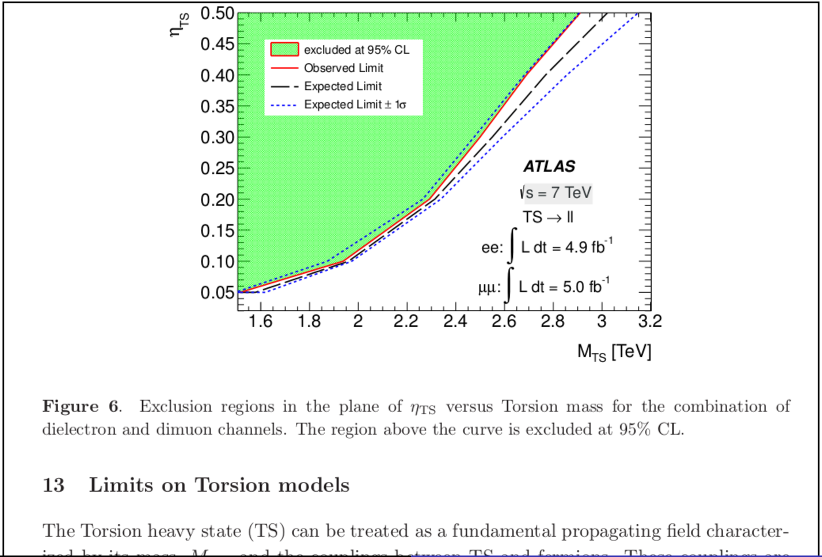


Figure 6. Exclusion regions in the plane of η_{TS} versus Torsion mass for the combination of dielectron and dimuon channels. The region above the curve is excluded at 95% CL.

13 Limits on Torsion models

The Torsion heavy state (TS) can be treated as a fundamental propagating field characterized by its mass M_{TS} and the couplings between TS and fermions. These couplings are

HepData entry describing a single figure

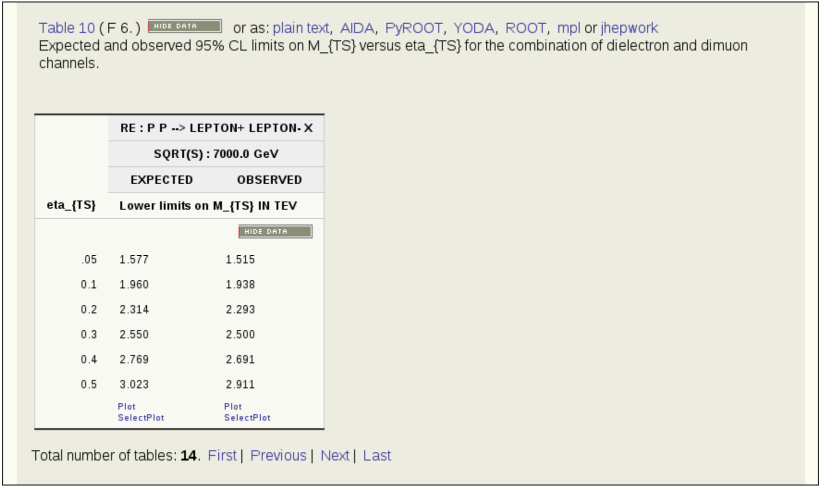


Table 10 (F 6.) HIDE DATA or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork
Expected and observed 95% CL limits on M_{TS} versus eta_{TS} for the combination of dielectron and dimuon channels.

	RE : P P --> LEPTON+ LEPTON- X	
	SQRT(S) : 7000.0 GeV	
	EXPECTED	OBSERVED
eta_{TS}	Lower limits on M_{TS} IN TEV	
		HIDE DATA
.05	1.577	1.515
0.1	1.960	1.938
0.2	2.314	2.293
0.3	2.550	2.500
0.4	2.769	2.691
0.5	3.023	2.911
	Plot SelectPlot	Plot SelectPlot

Total number of tables: **14**. First | Previous | Next | Last

Figure inside a publication and the corresponding HepData entry

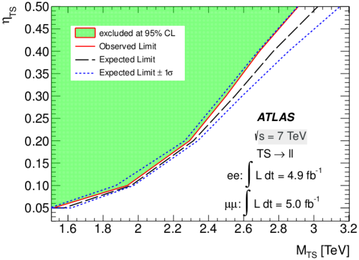


Figure 6. Exclusion regions in the plane of η_{TS} versus Torsion mass for the combination of dielectron and dimuon channels. The region above the curve is excluded at 95% CL.

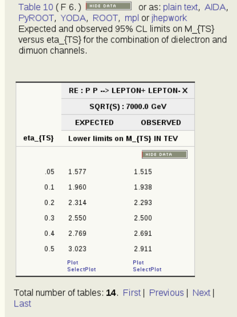


Table 10 (F 6.) or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork
Expected and observed 95% CL limits on M_{TS} versus eta_{TS} for the combination of dielectron and dimuon channels.

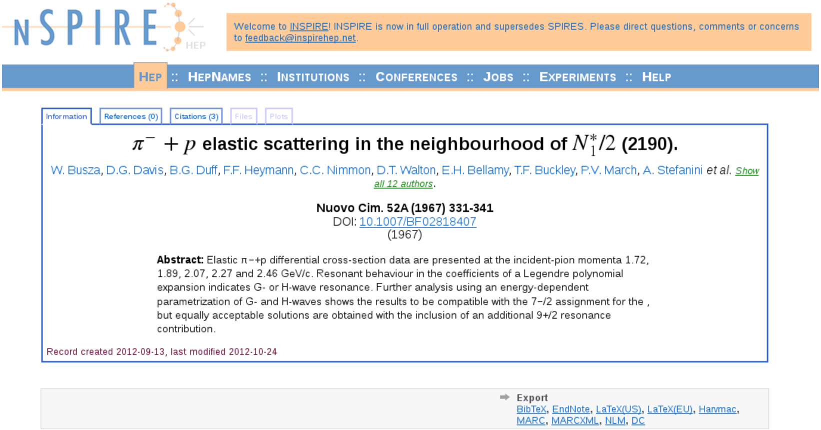
	RE : P P --> LEPTON+ LEPTON- X	
	SQRT(S) : 7000.0 GeV	
	EXPECTED	OBSERVED
eta_{TS}	Lower limits on M_{TS} IN TEV	
.05	1.577	1.515
0.1	1.960	1.938
0.2	2.314	2.293
0.3	2.550	2.500
0.4	2.769	2.691
0.5	3.023	2.911
	Plot SelectPlot	Plot SelectPlot

Total number of tables: **14**. First | Previous | Next | Last

Why should we integrate?

- INSPIRE contains publications and high-quality meta-data, HepData contains data.
- HepData stores only metadata of datasets and very minimal description of publications.
- Assignment of Digital Object Identifiers.

A bibliographic record before...



Welcome to INSPIRE! INSPIRE is now in full operation and supersedes SPIRES. Please direct questions, comments or concerns to feedback@inspirehep.net.

HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: HELP
Information
References (0)
Citations (3)
Files
Plots

$\pi^- + p$ elastic scattering in the neighbourhood of $N_1^*/2$ (2190).

W. Busza, D.G. Davis, B.G. Duff, F.F. Heymann, C.C. Nimmon, D.T. Walton, E.H. Bellamy, T.F. Buckley, P.V. March, A. Stefanini *et al.* *Show all 12 authors*.

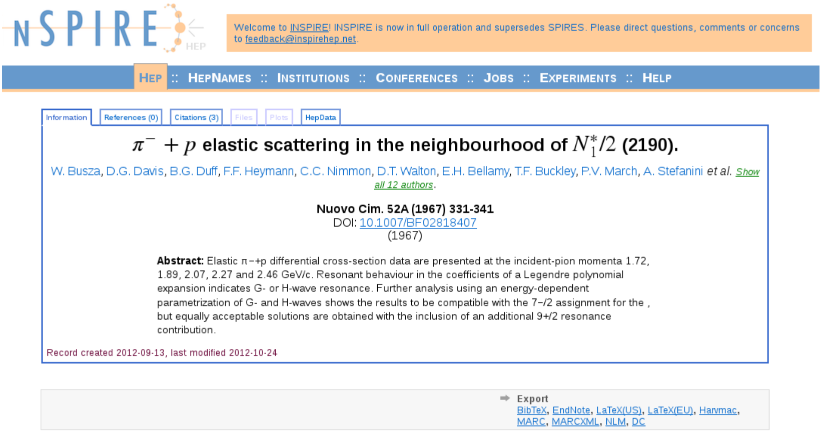
Nuovo Cim. 52A (1967) 331-341
DOI: 10.1007/BF02818407
(1967)

Abstract: Elastic π−+p differential cross-section data are presented at the incident-pion momenta 1.72, 1.89, 2.07, 2.27 and 2.46 GeV/c. Resonant behaviour in the coefficients of a Legendre polynomial expansion indicates G- or H-wave resonance. Further analysis using an energy-dependent parametrization of G- and H-waves shows the results to be compatible with the 7−/2 assignment for the , but equally acceptable solutions are obtained with the inclusion of an additional 9+/2 resonance contribution.

Record created 2012-09-13, last modified 2012-10-24

Export
BibTeX, EndNote, LaTeX(US), LaTeX(EU), Harvmac, MARC, MARCXML, NLM, DC

... and after


INSPIRE HEP
HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: HELP
Information
References (0)
Citations (3)
Files
Plots
HepData

Welcome to INSPIRE! INSPIRE is now in full operation and supersedes SPIRES. Please direct questions, comments or concerns to feedback@inspirehep.net.

$\pi^- + p$ elastic scattering in the neighbourhood of $N_1^*/2$ (2190).

W. Busza, D.G. Davis, B.G. Duff, F.F. Heymann, C.C. Nimmon, D.T. Walton, E.H. Bellamy, T.F. Buckley, P.V. March, A. Stefanini *et al.* *Show all 12 authors*.

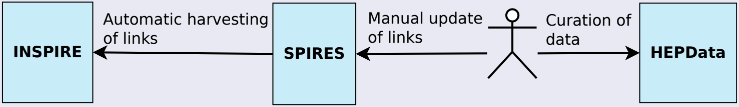
Nuovo Cim. 52A (1967) 331-341
DOI: 10.1007/BF02818407
(1967)

Abstract: Elastic π−+p differential cross-section data are presented at the incident-pion momenta 1.72, 1.89, 2.07, 2.27 and 2.46 GeV/c. Resonant behaviour in the coefficients of a Legendre polynomial expansion indicates G- or H-wave resonance. Further analysis using an energy-dependent parametrization of G- and H-waves shows the results to be compatible with the 7−/2 assignment for the , but equally acceptable solutions are obtained with the inclusion of an additional 9+/2 resonance contribution.

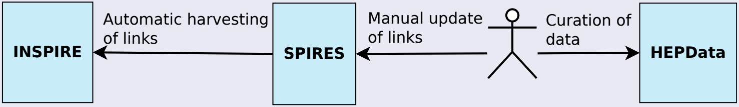
Record created 2012-09-13, last modified 2012-10-24

Export
BibTeX, EndNote, LaTeX(US), LaTeX(EU), Harvmac, MARC, MARCXML, NLM, DC

Interaction between systems before the integration



Interaction between systems before the integration



INSPIRE

Automatic harvesting of links

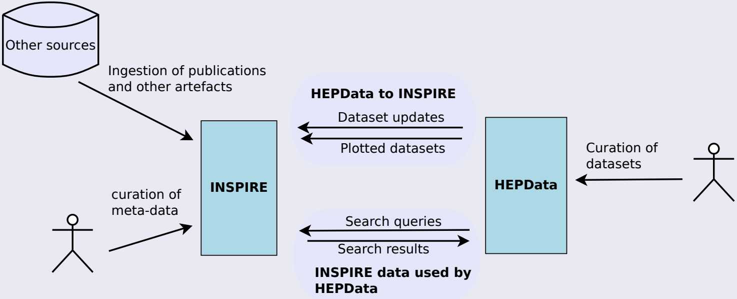
SPIRES

Manual update of links

Curation of data

HEPData

… and after the integration



Other sources

Ingestion of publications and other artefacts

curation of meta-data

INSPIRE

HEPData to INSPIRE

Dataset updates

Plotted datasets

Search queries

Search results

INSPIRE data used by HEPData

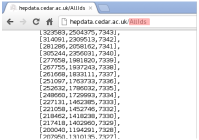
HEPData

Curation of datasets

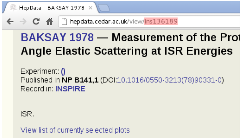
Harvesting

Assumption: We did not want to make deep modifications to HepData

- HepData does not maintain dates nor times of the last modification of a dataset
- Extensions of HepData:
    - HepData exports all the INSPIRE identifiers of records for which datasets are stored
    - HepData allows to access datasets by addressing them with INSPIRE id and number of a dataset



```
[323583,2504375,7343],
[314091,2309513,7342],
[281286,2058162,7341],
[305244,2356031,7340],
[277658,1981820,7339],
[267755,1937243,7338],
[261668,1833111,7337],
[251097,1763733,7336],
[252632,1786032,7335],
[248660,1729993,7334],
[227131,1462385,7333],
[221058,1452746,7332],
[218462,1418238,7330],
[217418,1402960,7329],
[200040,1194291,7328],
[207950,1310135,7327],
```



BAKSAY 1978 — Measurement of the Prot
Angle Elastic Scattering at ISR Energies

Experiment: ()
Published in NP B141,1 (DOI:10.1016/0550-3213(78)90331-0)
Record in: INSPIRE

ISR.

View list of currently selected plots

Harvesting (2)

- INSPIRE reads the entire list of publication IDs stored in HD and retrieves all datasets related to those publications
- Datasets within a single publications are matched with existing INSPIRE records and necessary updates are applied

Storing datasets in INSPIRE

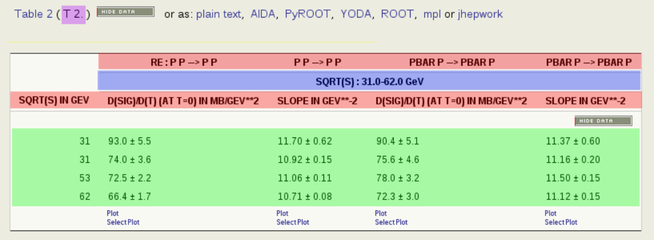


Table 2 (T 2.) HIDE DATA or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork

	RE : P P --> P P	P P --> P P	PBAR P --> PBAR P	PBAR P --> PBAR P
	SQRT(S) : 31.0-62.0 GeV			
SQRT(S) IN GEV	D(SIG)/D(T) (AT T=0) IN MB/GEV**2	SLOPE IN GEV**-2	D(SIG)/D(T) (AT T=0) IN MB/GEV**2	SLOPE IN GEV**-2
				HIDE DATA
31	93.0 ± 5.5	11.70 ± 0.62	90.4 ± 5.1	11.37 ± 0.60
31	74.0 ± 3.6	10.92 ± 0.15	75.6 ± 4.6	11.16 ± 0.20
53	72.5 ± 2.2	11.06 ± 0.11	78.0 ± 3.2	11.50 ± 0.15
62	66.4 ± 1.7	10.71 ± 0.08	72.3 ± 3.0	11.12 ± 0.15
	Plot Select Plot	Plot Select Plot	Plot Select Plot	Plot Select Plot

- 🟦 Data qualifiers (MARC: 6531)
- 🟩 Data (Attached data file, not stored in MARC - not metadata)
- 🟥 Column headers and titles (MARC: 910)
- 🟪 Part of a reference to a paper (MARC: 786)

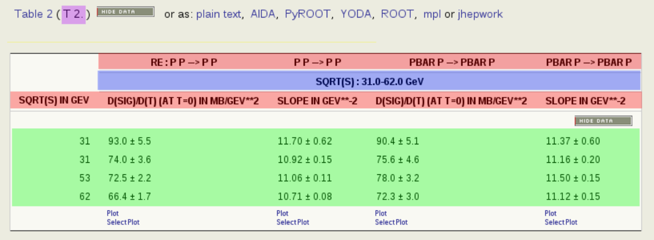


Table 2 (T 2.) HIDE DATA or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork

	RE : P P --> P P		PBAR P --> PBAR P	
		P P --> P P		PBAR P --> PBAR P
	SQRT(S) : 31.0-62.0 GeV			
SQRT(S) IN GEV	D(SIG)/D(T) (AT T=0) IN MB/GEV**2	SLOPE IN GEV**-2	D(SIG)/D(T) (AT T=0) IN MB/GEV**2	SLOPE IN GEV**-2
				HIDE DATA
31	93.0 ± 5.5	11.70 ± 0.62	90.4 ± 5.1	11.37 ± 0.60
31	74.0 ± 3.6	10.92 ± 0.15	75.6 ± 4.6	11.16 ± 0.20
53	72.5 ± 2.2	11.06 ± 0.11	78.0 ± 3.2	11.50 ± 0.15
62	66.4 ± 1.7	10.71 ± 0.08	72.3 ± 3.0	11.12 ± 0.15
	Plot Select Plot	Plot Select Plot	Plot Select Plot	Plot Select Plot

```
001__ 1157867
245__ $$9HEPDATA$$aAdditional data from: A MEASUREMENT OF anti-p p AND p p ELASTIC (...)
336__ $$tDATASET
520__ $$9HEPDATA
6531_ $$c4$$c3$$c2$$c1$$kSQRT(S)$$v31.0-62.0 GeV
710__ $$gAMES-BOLOGNA-CERN-DORTMUND-HEIDELBERG-WARSAW COLLABORATION
786__ $$hT 2.$$q2$$rCERN-EP/84-105$$w204422
8564_ $$uhttp://inspirehep.net/record/1157867/files/Data.txt$$ydata extracted from the table
910__ $$dSQRT(S) IN GEV$$n0
910__ $$dD(SIG)/D(T) (AT T=0) IN MB/GEV**2$$n1$$tRE : P P --&gt; P P
910__ $$dSLOPE IN GEV**-2$$n2$$tP P --&gt; P P
910__ $$dD(SIG)/D(T) (AT T=0) IN MB/GEV**2$$n3$$tPBAR P --&gt; PBAR P
910__ $$dSLOPE IN GEV**-2$$n4$$tPBAR P --&gt; PBAR P
911__ $$x1$$y4
980__ $$aDATA
```

Improvements in the display

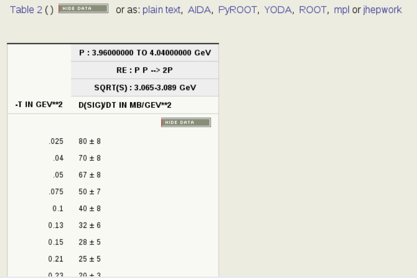
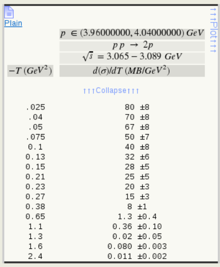


Table 2 ( ) or as: plain text, AIDA, PyROOT, YODA, ROOT, mpl or jhepwork

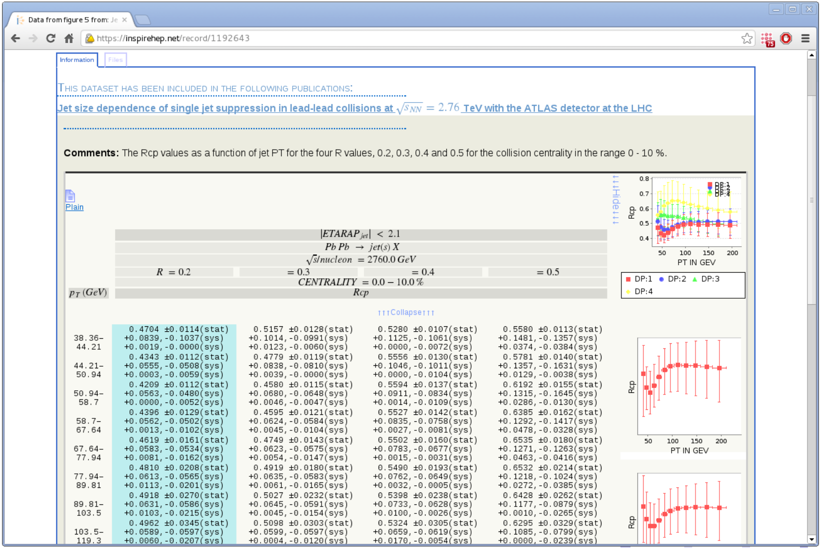
	P : 3.96000000 TO 4.04000000 GeV
	RE : P P --> 2P
	SQRT(S) : 3.065-3.089 GeV
-T IN GEV**2	D(SIG)/DT IN MB/GEV**2
.025	80 ± 8
.04	70 ± 8
.05	67 ± 8
.075	50 ± 7
0.1	40 ± 8
0.13	32 ± 6
0.15	28 ± 5
0.21	25 ± 5
0.23	20 ± 3



	$p \in (3.96000000, 4.04000000)\ GeV$
	$p\,p \rightarrow 2p$
	$\sqrt{s} = 3.065 - 3.089\ GeV$
$-T\ (GeV^2)$	$d(\sigma)/dT\ (MB/GeV^2)$
.025	80 ±8
.04	70 ±8
.05	67 ±8
.075	50 ±7
0.1	40 ±8
0.13	32 ±6
0.15	28 ±5
0.21	25 ±5
0.23	20 ±3
0.27	15 ±3
0.38	8 ±1
0.65	1.3 ±0.4
1.1	0.36 ±0.10
1.3	0.02 ±0.05
1.6	0.080 ±0.003
2.4	0.011 ±0.002

- Data qualifiers parsed and displayed in a more intuitive way
- Directly integrated plotting facility
- More compact view

Plotting of data integrated within the dataset view



This dataset has been included in the following publications:

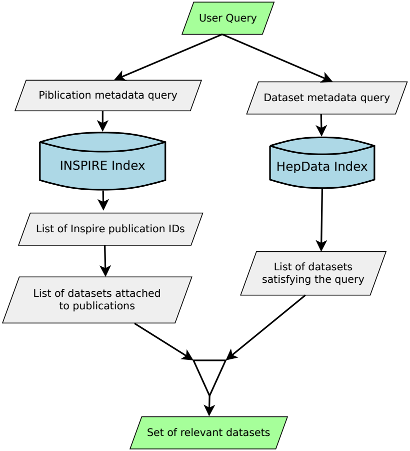
Jet size dependence of single jet suppression in lead-lead collisions at $\sqrt{s_{NN}} = 2.76$ TeV with the ATLAS detector at the LHC

Comments: The Rcp values as a function of jet PT for the four R values, 0.2, 0.3, 0.4 and 0.5 for the collision centrality in the range 0 - 10 %.

	$\|ETARAP_{jet}\| < 2.1$			
	$Pb\ Pb \rightarrow jet(s)\ X$			
	$\sqrt{s}/nucleon = 2760.0\ GeV$			
	$R = 0.2$	$= 0.3$	$= 0.4$	$= 0.5$
	$CENTRALITY = 0.0 - 10.0\ \%$			
$p_T\ (GeV)$	Rcp			
38.36–44.21	0.4704 ±0.0114(stat) +0.0839,-0.1037(sys) +0.0019,-0.0000(sys)	0.5157 ±0.0128(stat) +0.1014,-0.0991(sys) +0.0123,-0.0060(sys)	0.5280 ±0.0107(stat) +0.1125,-0.1061(sys) +0.0000,-0.0072(sys)	0.5580 ±0.0113(stat) +0.1481,-0.1357(sys) +0.0374,-0.0384(sys)
44.21–50.94	0.4343 ±0.0112(stat) +0.0555,-0.0508(sys) +0.0003,-0.0059(sys)	0.4779 ±0.0119(stat) +0.0838,-0.0810(sys) +0.0039,-0.0000(sys)	0.5556 ±0.0130(stat) +0.1046,-0.1011(sys) +0.0000,-0.0104(sys)	0.5781 ±0.0140(stat) +0.1357,-0.1631(sys) +0.0129,-0.0038(sys)
50.94–58.7	0.4209 ±0.0112(stat) +0.0563,-0.0480(sys) +0.0000,-0.0052(sys)	0.4580 ±0.0115(stat) +0.0680,-0.0648(sys) +0.0046,-0.0047(sys)	0.5594 ±0.0137(stat) +0.0911,-0.0834(sys) +0.0014,-0.0109(sys)	0.6192 ±0.0155(stat) +0.1315,-0.1645(sys) +0.0286,-0.0130(sys)
58.7–67.64	0.4396 ±0.0129(stat) +0.0562,-0.0502(sys) +0.0013,-0.0102(sys)	0.4595 ±0.0121(stat) +0.0624,-0.0584(sys) +0.0045,-0.0104(sys)	0.5527 ±0.0142(stat) +0.0835,-0.0758(sys) +0.0027,-0.0081(sys)	0.6385 ±0.0162(stat) +0.1292,-0.1417(sys) +0.0478,-0.0328(sys)
67.64–77.94	0.4619 ±0.0161(stat) +0.0583,-0.0534(sys) +0.0081,-0.0162(sys)	0.4749 ±0.0143(stat) +0.0623,-0.0575(sys) +0.0054,-0.0147(sys)	0.5502 ±0.0160(stat) +0.0783,-0.0677(sys) +0.0015,-0.0031(sys)	0.6535 ±0.0180(stat) +0.1271,-0.1263(sys) +0.0463,-0.0416(sys)
77.94–89.81	0.4810 ±0.0208(stat) +0.0613,-0.0565(sys) +0.0113,-0.0201(sys)	0.4919 ±0.0180(stat) +0.0635,-0.0583(sys) +0.0061,-0.0165(sys)	0.5490 ±0.0193(stat) +0.0762,-0.0649(sys) +0.0032,-0.0005(sys)	0.6532 ±0.0214(stat) +0.1218,-0.1024(sys) +0.0272,-0.0385(sys)
89.81–103.5	0.4918 ±0.0270(stat) +0.0631,-0.0586(sys) +0.0103,-0.0215(sys)	0.5027 ±0.0232(stat) +0.0645,-0.0591(sys) +0.0045,-0.0154(sys)	0.5398 ±0.0238(stat) +0.0733,-0.0628(sys) +0.0100,-0.0026(sys)	0.6428 ±0.0262(stat) +0.1177,-0.0879(sys) +0.0010,-0.0265(sys)
103.5–119.3	0.4962 ±0.0345(stat) +0.0589,-0.0597(sys) +0.0060,-0.0207(sys)	0.5098 ±0.0303(stat) +0.0599,-0.0597(sys) +0.0004,-0.0120(sys)	0.5324 ±0.0305(stat) +0.0659,-0.0619(sys) +0.0170,-0.0054(sys)	0.6295 ±0.0329(stat) +0.1085,-0.0799(sys) +0.0000,-0.0239(sys)

HepData using INSPIRE search engine

- INSPIRE store meta-data about publications
- HepData stores only meta-data describing separate datasets
- Using publication-level attributes indexed in INSPIRE allows to increase the accuracy of HepData search


User Query
Piblication metadata query
Dataset metadata query
INSPIRE Index
HepData Index
List of Inspire publication IDs
List of datasets attached to publications
List of datasets satisfying the query
Set of relevant datasets

Conclusions

The integration betweem HepData and INSPIRE extends the possibilities of users of both systems and improves their experience. However, some problems have been identified:

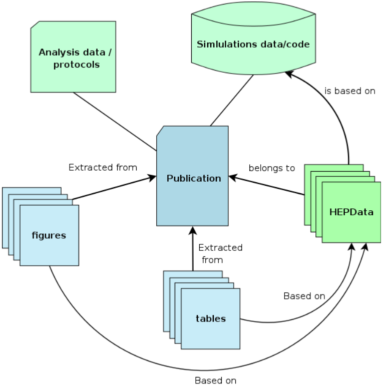
- Changes in the display format of HepData can have non-predictible effects on the INSPIRE dataset
- We need to harvest the entire database of HepData every time
- HepData is expected to provide the correct data, Inspire is expected to archive the publication data

Harvesting 2.0

- Data exported in fixed XML format
- All modification in HepData marked with timestamps
- Retrieving only publications for which at least one dataset has been updated
- Datasets should never be removed in INSPIRE.
- Small dataset changes (mistakes in transcription) : Update the record
- Severe updates (Data has been completely reuploaded) : Create a new record linking to the old one

Open Linked Data

- Linking datasets with figures (and other artefacts) in an automatic way
- Exposing data behind publications in RDF
- Assignment of Digital Object Identifiers



Questions?

piotr.praczyk@cern.ch