Results from the PARSE.Insight project

A survey on HEP data preservation, re-use and (open) access
http://arxiv.org/abs/0906.0485

• Background information
• The HEP Survey
• Conclusions

Andre Holzner (FP7/CERN), Peter Igo-Kemenes (Gjøvik/CERN), Salvatore Mele (CERN)
Actors

Alliance for Permanent Access

Group of stakeholders in digital preservation of data and publications. Purpose: best practices and dialogue with the EC

- Research (CERN, ESA, STFC, Helmholtz, MPG)
- National Libraries (Germany, Netherlands, UK)
- Publishers (STM association)

CERN involvement through Scientific Information Service/Open Access
e-Infrastructure = connectivity (GEANT) + Grid (EGEE)
Other activities = repositories (i.e. DB for publications)
Possible future = extension to data (2010-2013 + FP8)
PARSE.Insight

http://www.parse-insight.eu/

• FP7 project, 9 partners, 1.2M€, 3/2008–4/2010:
  • Science  (CERN, ESA, STFC, MPG)
  • Libraries (KB,DNB,..)
  • Publishers (STM)

• Interdisciplinary study on research-data preservation

• Detailed case studies: HEP (+Earth observation, +...)

• Deliverables to inform FP7/8 policy and strategy:
  – Threats/opportunities in data preservation today
  – Roadmap for the future

Salvatore Mele | CERN | Dec. 7, 2009
Just survey what HEP thinks of preservation...  
... but it aims to shed light on:

- Motivations vs. Concerns
- Threats vs. Opportunities
- Wishes vs. Obstacles

- To provide evidence to decision-making process
- To provide input to the DPHEP initiative
- To know what the “community” thinks
Survey strategy and response

- Livetime of 3 months 10/08-01/09
- Collaborations/theory mailing lists + SPIRES
- 1’200 answers (74% exp., 25% th.). Target ~20’000
Survey structure

1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
Survey demographics

Reflects the demographics the community

Which of the following best describes your position? (top/blue: theorists, bottom/green: experimentalists)

- PhD student: 19.4% (blue), 24.2% (green)
- Post-doctoral fellow: 26.0% (blue), 22.3% (green)
- Researcher (permanent position): 20.4% (blue), 30.7% (green)
- Professor: 34.2% (blue), 22.8% (green)
Survey demographics

Reflects the demographics of experiments

In which experiments are you / have you been involved?

- ALICE
- ATLAS
- CMS
- LHCb
- ALEPH, DELPHI, OPAL, L3
- BaBar, Belle
- CDF, D0
- H1, ZEUS
- CLEO
- STAR, PHENIX, BRAHMS, PHOBOS
- Neutrino experiments
- Kaon experiments
- Fixed target experiments
- ILC, SLHC or other future projects
- Other experiments

Salvatore Mele | CERN | Dec. 7, 2009
1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
The importance of preservation

In your opinion, how important is the issue of data preservation?
(top/blue: theorists, bottom/green: experimentalists)

- Irrelevant: 0.4%
- Moderately important: 3.3%
- Important: 15.2%
- Very important: 41.7%
- Crucial: 40.5%
The importance of preservation

In your opinion, how important is the issue of data preservation?

- Irrelevant: 0.6% (5 years in HEP), 1.0% (< 5 years in HEP)
- Moderately important: 3.1% (5 years in HEP), 8.9% (< 5 years in HEP)
- Important: 22.9% (5 years in HEP), 23.0% (< 5 years in HEP)
- Very important: 42.7% (5 years in HEP), 40.2% (< 5 years in HEP)
- Crucial: 30.2% (5 years in HEP), 27.4% (< 5 years in HEP)
The importance of preservation

Future independent checks

b) Preserved data would allow future independent verification of results
   (top/blue: theorists, bottom/green: experimentalists)

Combine with future data
c) Preserved data could be reused in combination with future data
   (top/blue: theorists, bottom/green: experimentalists)

Re-analyse for future theories
d) Preserved data could be re-analysed in the light of new theories
   or experimental results
   (top/blue: theorists, bottom/green: experimentalists)

Teaching and outreach
e) Preserved data could be used for teaching and outreach
   (top/blue: theorists, bottom/green: experimentalists)
Would preserved data enable better science?

Do you think that access to data from past experiments could have improved your scientific results?
(top/blue: theorists, bottom/green: experimentalists)

1. Strong argument to ask support for preservation
2. Cannot separate preservation, re-use and access
Re-use
(Open) Access
Preservation
Did anything go wrong so far?

Do you think that in the past important HEP data have been lost?
(top/blue: theorists, bottom/green: experimentalists)

- Yes: 42.1%
- No: 57.9%

Over optimistic? Over pessimistic?
Survey structure

1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
What to preserve?

At what level of detail should data be preserved?
(top/blue: theorists, bottom/green: experimentalists)

- **Publications ++**
  - Information from published tables and figures, e.g. numerical information in electronic form
  - Backup information which did not fit in your publication (e.g. additional numerical information, figures and tables, comparison of data and simulation)
  - Multi-dimensional distributions (differential cross sections in many variables, likelihood distributions) which cannot be fully detailed in the publication.

- **More table/figures**
  - Event-by-event higher-level objects (e.g. four-vectors), together with appropriate information allowing some re-analysis of the data.

- **Multi dimensional info**
  - Raw data (together with appropriate access and interpretation "tools") allowing complete re-analysis of the data.

- **Four vectors**
  - 62.3%
  - 74.9%

- **Raw data**
  - 50.7%
  - 61.6%

- **Multi dimensional info**
  - 58.7%
  - 60.4%

- **Four vectors**
  - 57.0%
  - 69.7%

- **Publications ++**
  - 43.9%
  - 45.5%
Survey structure

1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
When would data be available for preservation?

- After data taking: 6.8%
- After publication: 31.9%
- Additional delay: 21.4%
- End of experiment: 39.3%
- Never: 0.6%
When would data be available for preservation?

- After data taking
- After publication
- Additional delay
- End of experiment
- Never

- > 5 years in HEP
- < 5 years in HEP

Salvatore Mele | CERN | Dec. 7, 2009
Where to preserve?

Where should data be preserved?
(top/blue: theorists, bottom/green: experimentalists)

- On a site connected to the experiment/laboratory: Labs/Exp’ts
  - Theorists: 41.2%
  - Experimentalists: 61.0%

- On a "neutral" platform, an infrastructure such as ADS, arXiv, CDS or SPIRES adapted to house data: Something like arXiv/SPIRES
  - Theorists: 85.6%
  - Experimentalists: 62.8%

- On a platform managed by a journal publisher: Something like journals
  - Theorists: 3.7%
  - Experimentalists: 3.6%
Survey structure

1. Demographics
2. The importance of preservation
3. What to preserve
4. When, how and where to preserve
5. Threats
Reality check #1: how though is it to preserve?

How much additional effort do you think is needed for the preservation of your data in a re-usable form (in percent of the overall effort invested in the production and analysis of the data)?

- LEP: 6.7%, 6.8%, 5.3%, 6.0%
- CDF/D0: 47.9%, 50.7%, 46.7%, 50.7%
- H1/ZEUS: 37.0%, 32.4%, 42.7%, 36.9%
- LHC: 8.4%, 10.1%, 5.3%, 6.4%

As a percentage of the data-taking effort

Salvatore Mele | CERN | Dec. 7, 2009
Reality check #2: when to start?

In your opinion, when should this effort start in order to be the most effective?

- Before data taking:
  - LEP: 17.9%
  - CDF/D0: 20.1%
  - H1/ZEUS: 23.0%
  - LHC: 30.6%

- Concurrently to data taking:
  - LEP: 42.7%
  - CDF/D0: 57.6%
  - H1/ZEUS: 50.0%
  - LHC: 42.4%

- After data taking:
  - LEP: 39.3%
  - CDF/D0: 22.2%
  - H1/ZEUS: 27.0%
  - LHC: 27.0%

....which means now!
Reality check #3: is it doable?

Will your experiment/collaboration/organisation be able to invest this effort?

Yes
- LEP: 10.9%
- CDF/D0: 7.4%
- H1/ZEUS: 22.7%
- LHC: 14.6%

No
- LEP: 84.9%
- CDF/D0: 84.6%
- H1/ZEUS: 9.3%
- LHC: 6.4%

Don't know
- LEP: 4.2%
- CDF/D0: 8.1%
- H1/ZEUS: 68.0%
- LHC: 79.0%
Ideal-case worries: getting credit

To what extent are you concerned about the following issues related to giving access to preserved data?

a) Preserved data could be used without giving proper credit to the original authors
   (top/blue: theorists, bottom/green: experimentalists)

- Not concerned
  - Theorists: 25.6%
  - Experimentalists: 14.0%
- Moderately concerned
  - Theorists: 27.4%
  - Experimentalists: 28.3%
- Concerned
  - Theorists: 27.4%
  - Experimentalists: 32.2%
- Very concerned
  - Theorists: 14.0%
  - Experimentalists: 18.0%
- Gravely concerned
  - Theorists: 5.6%
  - Experimentalists: 7.5%
Ideal-case worries: inflation/noise

To what extent are you concerned about the following issues related to giving access to preserved data?

b) Uncontrolled access to data may lead to an inflation of incorrect results (top/blue: theorists, bottom/green: experimentalists)

- Not concerned: 21.0% (theorists), 7.0% (experimentalists)
- Moderately concerned: 26.6% (theorists), 14.7% (experimentalists)
- Concerned: 23.4% (theorists), 26.9% (experimentalists)
- Very concerned: 19.6% (theorists), 31.4% (experimentalists)
- Gravely concerned: 9.3% (theorists), 19.9% (experimentalists)
Ideal-case worries: documentation

If you were to re-use preserved data, to what extent would you be concerned by the following scenarios?

d) I am not using the data correctly
(top/blue: theorists, bottom/green: experimentalists)

- Not concerned: 7.4%, 5.4%
- Moderately concerned: 14.8%, 11.1%
- Concerned: 31.0%, 28.3%
- Very concerned: 30.0%, 34.6%
- Gravely concerned: 16.7%, 20.6%

Salvatore Mele | CERN | Dec. 7, 2009
Conclusions

Provided quantitative insight on needs and opportunities in data preservation in HEP

• Strong awareness/support of necessity/urgency of data preservation

• Consensus across experiments and with theory

• Suggestion that global initiatives can be the solution
Thank you!

Salvatore.Mele@cern.ch

More details on PARSE.Insight HEP survey:
http://arxiv.org/abs/0906.0485