briefing paper

digital preservation europe

Data Preservation, Reuse and (Open) Access in High-Energy Physics

High-Energy Physics is a discipline relying on scientific instruments of unprecedented size and complexity, yielding a "deluge" of non-reproducible data. Surprisingly, preservation, reuse and open access to these data, which are deeply intertwined, are not high on the community agenda. Their inception, implementation and ultimate success are under siege from issues commonly found in the areas of digital preservation. This brief paper gives an introduction to this emerging debate

High-Energy Physics (HEP) aims to understand the laws of physics governing the Universe by investigating the fundamental constituents of nature (elementary particles) and their interactions (forces).

HEP research is epitomized by gigantic experimental facilities, accelerators and detectors, which exceed by far the financial capabilities of single countries. Experimentation is therefore concentrated at transnational research centers such as CERN in Europe, FERMILAB in the US, and a few others. Theoretical HEP scientists are spread over hundreds of institutions worldwide but they are interacting in a vibrant trans-national network. The HEP community is relatively small and tightly-knit. It has a dense social network and manages its own scholarly communication platforms.

CERN's Large Hadron Collider (LHC) constitutes the most complex scientific tool ever built. This "flagship" of European science will be able to explain, among other basic issues, the appearance of matter in the universe and perhaps the nature of "dark matter" and "dark energy" which make up the largest part of our universe. The LHC will produce collisions between protons 40 million times a second which are observed by detectors of the size of five-storey buildings, crammed with electronic sensors: think of a gigantic 100-Mega-Pixel digital camera taking 40 million pictures per second, producing tens of Petabytes of data per year. Even though interesting data are filtered on the fly, the LHC will place HEP at the front of the "data deluge" of modern e-Science. GRID computing has been developed precisely for the processing of this amount of data, the preservation of which is a burning concern.

HEP has a long record of pioneering scholarly communication [1]. About four decades ago HEP scientists have introduced the ante-litteram Open Access paradigm by disseminating paper-based "preprints" to their peers through ordinary mail; free electronic metadata-bases, such as Stanford's SLAC SPIRES and the archetypal full-text repository, arXiv.org, followed soon; the Web was invented at CERN to facilitate communication between peers worldwide. Against this background, it may come as a surprise that the HEP community is not leading in matters of data preservation and reuse.
An immediate explanation derives from the sheer size and complexity of the HEP primary data; however, sociological and financial issues also play a role. Academic pressure to produce immediate scientific results assigns a low priority to the issues of preservation, and the personal and material costs involved are not budgeted in the running of large-scale facilities.

The goal of preservation is the ability of reusing the data at a later stage. This will only work if the ancillary information describing the data and the experimental conditions are properly stored along with the data, which has to happen practically on-line since the required expert knowledge evaporates rapidly.

Data preservation within HEP should be carried out with four different continua in mind: (1) researchers who created the data and would like to re-analyze them after their experimental facility has been closed down, typically within a decade; (2) researchers working on similar experiments who would like to access the data, within one year or less, to analyze them together with their own data; (3) future scientists dealing with similar subjects who would like to access the preserved data for comparison with theirs or even for a joint analysis, on a timescale comparable to current HEP projects, one or two decades; (4) theoretical physicists wanting to re-interpret the preserved data in the light of new insights or to use them for testing new models, on a timescale from months, to years, to decades.

The ability to re-use preserved data opens a Pandora's box of issues that have never been addressed within the HEP community such as the data ownership, authenticity, (open) access, credit, accountability, reproducibility, as well as novel approaches required to peer-reviewing publications based on preserved data.

The PARSE.Insight project supported by the European Commission under FP7 aims to provide insight in issues of preservation to the records of science by running broad inter-disciplinary surveys and specific case studies. One of these is targeting the HEP community. Within one month on-line, the HEP survey collected over one thousand answers (about 5% of all active scientists in the field). A first analysis shows the following salient features:

• More than 90% of respondents think that the issue of data preservation is "important", "very important" or "crucial" for the benefit of the field;
• Among the main motivations for preservation, the following stand out as "very important" to "crucial": independent verification of scientific results (60%); combination of past and future data (60%); re-analysis in the light of new theories and future results (75%);
• About 45% of the respondents feel that access to the data from past experiments could have improved their scientific results and, worryingly, about 40% think that important HEP data have been lost in the past.

Concerning the investment of personal effort, the willingness seems overwhelming. Regarding the level of detail, 80% would be willing to provide numerical data in support of published tables and figures while 45% would go as far as to provide "raw" measurement data packaged for later reuse. However, 50% consider that the supplementary cost involved would be high, up to 10% of the effort required for data-taking and analysis or even more. This is a source of concern.

As a possible solution for the preservation of the records of HEP, 75% think that a trans-national infrastructure should be built. The most-recurrent qualifications are: reliability, stability of operation and funding; functionalities similar to Grid computing; unrestricted open access; neutrality and impartiality in data selection; the ability to cope with legacy software; a large international consensus on attributes, governance and operation.

In conclusion, there is no concerted effort today to preserve the data created by the costly and non-reproducible HEP facilities. This is due to the amount and complexity of the data and the lack of financial and academic incentives. Preliminary results from the PARSE.Insight project indicate however a strong commitment of the HEP community towards data preservation claiming a trans-national infrastructure. HEP may be regarded as a "worst case scenario" for which novel practices and solutions for preservation have to be worked out. In the track-record of the Web and the Grid, these could inspire other fields of science.

Notes

[1] R. Heuer, A. Holtkamp and S. Mele, Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics, Inform. Serv. Use 28 (2008) 83-96, arXiv:0805.2739

[2] P. A. Kreitz and T. C. Brooks, Sci. Tech. Libraries 24 (2003) 153, arXiv:physics/0309027; http://slac.stanford.edu/spires/hep/

[3] P. Ginsparg, Computers in Physics 8 (1994) 390; http://arxiv.org

[4] http://www.parse-insight.eu

[5] S. Mele, Preservation, re-use and (open) access in high energy physics data (or lack thereof), talk at the Third Annual WePreserve Conference, Nice, France, October 28-30, 2008 http://www.digitalpreservationeurope.eu/registries/materials/?author%5B%5D=40

Andre Holzner (1), **Peter Igo-Kemenes** (1, 2), **Salvatore Mele** (1)
(*1*) ***CERN**, European Organization for Nuclear Research, Geneva, Switzerland*
(*2*) ***Gjøvik University**, Gjøvik, Norway*