SLAC–PUB–13558
Mar 1, 2009

The Role of INSPIRE in HEP Data Preservation Efforts

Travis C. Brooks
SLAC, Stanford Linear Accelerator Center Library/SPIRES Databases[1]
Stanford University, Stanford, CA 94309, USA

Abstract INSPIRE is a new community resource for HEP literature and associated information. It is based on the combination of SPIRES' content and features and the powerful Invenio software developed at CERN. The INSPIRE service will come online in fall of 2009, and be run by CERN, DESY, Fermilab and SLAC. Data preservation, to be successful, must not only preserve the data, but must also organize it and allow it to be found by those who would make use of it, and resources such as INSPIRE are ideally positioned and ready to provide this organization and context. In addition, INSPIRE will soon be ready to provide storage of smaller datasets, such as high-level analysis objects, as stand-alone objects placed in the repository or as objects associated with an analysis paper. This small project could pave the way towards the context and organization which is one piece of the infrastructure needed for all levels of data preservation.

1 What is INSPIRE?

Researchers in HEP have long benefited from having the entire corpus of their literature at their fingertips. What began as a system of preprint dissemination among colleagues at large institutions and labs has evolved over time to become a comprehensive community resource for literature and associated information [?, ?]. SPIRES [?], run by SLAC, Fermilab, and DESY, has been operating for over 35 years, with a strong history of continuous innovation. Not only was SPIRES the first database on the web in 1991 [?], but it has consistently provided services to the community as (and sometimes before) they were demanded by the users [?]. CERN has also played a role in online dissemination of preprints in HEP, via CDS [?, ?], and KEK has provided scans of published papers for many years [?]. HEP's major laboratories have always been heavily involved in providing the infrastructure to communicate research results within the community, and these community-based resources have proven to be quite successful [?].

As SPIRES helped usher in the era of the World Wide Web it also presaged the growth of community driven, social networking information resources. For many HEP researchers, using SPIRES is a fundamental component of the research process, and this centrality within the field gives SPIRES the ability to provide other valuable services, such as a Jobs database. By responding to community needs and providing a central site for HEP research information, SPIRES provides more than just literature for the field. However as software grows more powerful, and as fields broaden to encompass more multidisciplinary work, we have identified opportunity, and the need, to build a system that continues to satisfy HEP researchers' needs for rapid unfettered communication, but that incorporates more sophisticated methods for

[1] Work partly supported by Department of Energy contract DE–AC02–76SF00515

building and maintaining community knowledge, and that builds connections to and from other disciplines.

This transition is the optimal time to take advantage of more modern understanding of how to provide useful functions like search, community networking, as well as the storage of associated objects, such as data. A workshop at SLAC in 2007 of the relevant HEP information providers (publishers, arXiv, NASA-ADS, PDG, Google Scholar and several others) addressed this opportunity [?]. Out of this workshop emerged INSPIRE, a joint project of CERN, DESY, Fermilab and SLAC to design and build the next-generation HEP literature service using SPIRES as a foundational framework [?]. The resulting system will blend the content and functionality of SPIRES with the powerful digital library software, Invenio [?], developed at CERN, and adapted for this purpose. The INSPIRE project has been ongoing for 1.5 years, has met the first two project milestones and is on schedule to deliver the initial service in Fall 2009. The project will continue through the end of 2010 with an intense effort to fully develop the functionality made possible by the new framework. The INSPIRE service will be, like SPIRES, a global service run by a collaboration of HEP labs: CERN, DESY, Fermilab and SLAC.

INSPIRE, as the successor to SPIRES, will keep HEP at the forefront of scholarly communication by combining modern technology and methodology with the experience of SPIRES and the HEP community. INSPIRE will provide the HEP community with a trusted central resource that meets the needs of today yet is prepared to grow with the needs of of the community tomorrow.

2 The Importance of Organization

There are many challenges to preserving data in HEP, and many of those are beyond the scope of the sort of lightweight web services that SPIRES and INSPIRE currently can provide. However, these sorts of resource can provider a crucial element to Data Preservation that is often overlooked when focusing on the more technical challenges inherent in preserving Petabytes of raw or reconstrcuted data in usable form. If HEP data is preserved, in any form, at any scale, it must be made accessible to (some part of) the community on a long term basis, and it must be findable. Access, and more crucially, findability, is central to preservation not only because is it impossible to use data which you cannot find, but also because continual use is one of the simplest guarantors of preservation. A central, community resource like INSPIRE is ideal to provide organisation and compilation of preserved data.

These requirements are non-trivial to satisfy over any appreciable amount of time. One might imagine a coalition of large experiments creating an infrastructure for preserving their data, but it is clear that experiments themselves are poorly suited to the task of preserving the data long term. While a large experiment might be able to keep a website with data archives available 10 years after the disbanding of the collaboration, via cooperation from the hosting lab, it would be difficult and uncertain. For smaller experiments, this sort of long term preservation would be nearly impossible on their own.

Smaller experiments make up the so-called ’long tail’ of HEP data. Of the 580 papers published in hep-ex in 2007 which SPIRES identified as comging from a collaboration, there were nearly 100 different collaborations listed. Most papers came from the large experiments running at that time (CDF, BaBar, DO, etc.), and these experiments were publishing around

50 papers a year, but there were over 100 papers from collaborations that had less than 10 papers. The importance of the large experiments as an influence on this long tail should not be underestimated. Large experiments not only make things slightly simpler by creating and utilizing central repositories, they also set best practices that enable and encourage smaller experiments to take advantage of the central resources, thus preserving data that would almost certainly be lost otherwise.

The natural organization of data is by experiment, and, at a finer-grained level, by analysis. This maps very nicely to the sort of data that is stored in literature databases like SPIRES. Creating links from literature to data on which an analysis paper is based not only enhances the information about the paper, but also provides a context for the data, and a social and technical infrastructure in which the data will be used and hence preserved. By far the best way to ensure preservation over time is to enable use. Things that are heavily used are preserved by neccessity.

Astrophysics provides such a linkage between papers and data in the NASA-ADS system [?], which links large data repositories such as CDS [?] and ESO [?] to the analysis papers based on that data, and the instrumentation papers that describe the data-taking apparatus. Many of these data repositories include access controls and allow control that can be tuned according the policy of the data depositor, their funding agency, or other interested parties. While HEP has a different culture and different needs than astrophysics, the basic infrastructure created in that field is relevant to HEP as a proof of principle and a rough roadmap. Astrophysics' use of data underscores the need for centralized repositories, and connections with everyday uses, like literature, that promote use and visibility of preserved data.

How long should the data be preserved, and how much utility does it provide? That is certainly a question for experimentalists, theorists, funding agencies and managers, not repositories. However as a literature repository, SPIRES can help answer the question of how often older data is used. For example, looking at the publication history of the four LEP experiments (ALEPH, DELPHI, L3 and OPAL) as recorded by SPIRES, we see that there has been quite some publication activity since the shutdown of LEP.

Figure 1: SPIRES Data for the number of papers and number of full collaboration publications for all 4 LEP experiments.

The majority of the activity was during the latter part of the run and just after, and the period of time from 2003 through 2008 represents the decline of the literature output of the collaborations. Yet even during this time, over 250 papers were written by the collaborations themselves.

In a future in which data objects are linked or stored with papers, or as objects in a similar location to papers, one can imagine that this sort of question would be much easier to answer. Tracking the use, by download or by citation, of data stored in central repositories and data that is citeable like papers would be rather trivial for repositories like INSPIRE. Providing citeable records for data objects, which would then naturally link to associated experiments, analysis papers and instrumentation papers, would be particularly easy to arrange, and would maximally integrate the data objects with the other information in the field.

3 INSPIRE and Smaller Data Objects

How can INSPIRE aid the Data preservation efforts in HEP? One salient feature of INSPIRE is the ability to deposit objects associated with a paper in the system. While Petabytes of raw data are probably not appropriate for this system as it currently exists, small to mid-size data objects that are likely to be useful for later analysis and archival purposes will fit very well in this system. INSPIRE is expected to come online in Fall 2009, and could immediately begin storing, for example, ROOT files associated with the analysis in a given paper. Analysis code and other high-level, many Mb size objects could be associated with papers and/or experiments, and thus deposited in INSPIRE, a third-party, community-maintained repository. They could be linked directly with a given paper, or created as independent objects with their own author lists and citations, and linked to the relevant papers.

Either at the time of release (fall 2009), or within a few months thereafter, INSPIRE would have all the tools needed to allow an experimental collaboration, or an individual author, to deposit a file of tens to hundreds of Mb to be associated with a paper and available for use and download by the general public. SPIRES, and thus INSPIRE, works closely with the reactions database group at Durham U. [**?**], and with arXiv.org, so that the submission of numerical data, data objects, and the paper itself, could conceivably be streamlined into one process with some little effort in the time after release of INSPIRE. Other functionality that could be added with small amounts of effort would be access control for the associated objects, with perhaps a predetermined set of policies regarding access.

The preservation of these high-level data objects is a small task, technically, when compared to the preservation of data objects much closer to the raw or reconstructed data, as discussed in much of the rest of this proceedings. However, this project can be accomplished readily in the next year or so, and can lay the foundation for an infrastructure for organizing and preserving the more complex, lower-level datasets.

However and wherever large datasets are preserved, the integration of smaller independent data objects in a community resource like INSPIRE will pave the way for connecting the larger datasets to the literature. INSPIRE can place datasets in their appropriate context, maximizing their utility to the community, which in turns helps guarantee their preservation, through reuse.

References

[1] L. Goldschmidt-Clermont, *Communication Patterns in High-Energy Physics*, 1965; published in High Energy Physics Libraries Webzine, issue 6, March 2002. http://library.cern.ch/HEPLW/6/papers/1/.

[2] L. Addis, http://www.slac.stanford.edu/spires/papers/history.html.

[3] http://www.slac.stanford.edu/spires.

[4] http://www.slac.stanford.edu/history/earlyweb/history.shtml.

[5] P. A. Kreitz and T. C. Brooks, Sci. Tech. Libraries **24** (2003) 153 [arXiv:physics/0309027].

[6] http://cds.cern.ch.

[7] C. O'Dell *et al.*, CERN-OPEN-2003-053
http://doc.cern.ch/archive/electronic/cern/preprints/open/open-2003-053.pdf.

[8] http://www-lib.kek.jp/KISS/kiss_prepri.html.

[9] A. Gentil-Beccot, S. Mele, A. Holtkamp, H. B. O'Connell and T. C. Brooks, J. Am. Soc. Inf. Sci. **60**, 150 (2009) [arXiv:0804.2701 [cs.DL]].

[10] http://indico.cern.ch/conferenceDisplay.py?confId=11611

[11] See press release at: http://www.interactions.org/cms/?pid=1026243

[12] http://cdsware.cern.ch.

[13] http://adswww.harvard.edu.

[14] http://cdsweb.u-strasbg.fr/

[15] http://www.eso.org/public/

[16] http://durpdg.dur.ac.uk/HEPDATA/REAC