Data And Software Preservation for Open Science (DASPOS)

For the past few years, the worldwide High Energy Physics (HEP) Community has been developing the background principles and foundations for a community-wide initiative to move in the direction of open access, preservation, and reuse of data collected and analyzed by the field. As a subcommittee of the International Committee on Future Accelerator (ICFA) the Study Group for Data Preservation in HEP (DPHEP) has held a number of meetings in different continents and published an initial report that is the most in-depth analysis of the issues produced to date.[1] Given the scope, breadth and depth of the sociological, technical and governance challenges there are many activities developing around this nucleus.

Recently, a laboratory consortium including CERN, DESY and CNRS in Europe submitted a proposal for International Long-term Data and Analysis Preservation (ILDAP) to develop and prototype a first set of data reuse archival and prototyping services in the context of the Large Hadron Collider (LHC) experiments. The ILDAP proposal represents an initial attempt to translate the recommendations and concepts outlined in the DPHEP report into a framework than could form the basis for Data and Analysis preservation in HEP. Because this is an initial foray into the technical aspects of curation in HEP, the ILDAP effort is largely a review and planning exercise. Its main focus is to establish the technical requirements for data preservation and access. It will explore, in conjunction with the HEP experiments, possible standards for common simplified data formats and metadata descriptors that could be adopted across the field. It will also enumerate and evaluate various data access and preservation technologies with the goal of preparing prototype systems for data validation and archiving by the end of the funding period.

A central focus of the ILDAP proposal is the coordination of efforts among the data stakeholders, namely the HEP experiments, host laboratories, and those interested in driving the technologies associated with the data/analysis curation efforts. In particular, there is a need to interface with a similar effort in the US. The proposal described here, Data and Software Preservation for Open Science (DASPOS), represents a first attempt to establish a formal collaboration tying together physicists from the CMS and ATLAS experiments at the LHC and the Tevatron experiments with experts in digital curation, heterogeneous high-throughput storage systems, large-scale computing systems, and grid access and infrastructure. Together, this group would represent the US counterpart to ILDAP, acting as a point of contact, a partner in dialogue, and a technological consort. These activities will be connected into DPHEP, the experimental collaborations, and other related projects in Europe, Asia and the USA. The intent is to define and execute a small set of initial, well-defined, small-scale activities to provide beneficial outcomes upon which a larger scale, longer term program can later be based.

A primary goal of the DASPOS effort would be to enable, toward the end of the funding period, a "**Curation Challenge**" where, for example, an ATLAS physicist might perform an analysis on curated and archived CMS data. The choice of a relatively narrow focus limits the scope of the proposal to the eventual demonstration of a targeted set of technologies, and is commensurate with the small size of the team available. The longitudinal nature of this effort, however, will allow experience and evaluation of the issues and solutions associated with a full example of data curation and access. The problems of technical achievement, policy, coordination and communication that arise will be relevant to any broader efforts in this domain. This small-scale test bed can thus serve as a microcosm for the global data curation efforts.

The DASPOS research and planning efforts would then be organized with the goal of a Curation Challenge as the focal point of interest. Before undertaking a full design of the prototype software and data archive, many issues will need to be resolved in conjunction with the ILDAP effort, the HEP experiments, and our colleagues from other disciplines. These include the following considerations, most of which mirror those outlined in the ILDAP proposal:

Establishment of Use Cases for Archived Data and Software. Taking the Curation Challenge as a basis, what data and software would need to be preserved to enable a physicist from a different experiment, for example, to complete a full-fledged analysis using another experiment's data? A survey of use cases should be the overriding design principle for the archive architecture, since the use of the data determines almost everything about how it will be curated, from the amount of mirroring required, when the data is archived, what connectivity to the storage is required, how the data is registered and retrieved, etc. These determinations may be different for different use cases; choices of optimization and the degree to which the

[1] DPHEP-2009-00, http://arxiv.org/pdf/0912.0255

infrastructure can be generalized will be a central result of these discussions. In addition, a survey of use cases leads naturally to an enumeration of what policies in terms of access, authorization, etc., will be required to implement distributed data access and re-use at some later time.

Survey of Commonality with other Disciplines. While the focus of this preliminary effort is centered on the HEP community, many other disciplines face similar issues for the curation of their data, software, and documentation. A "generic" archival platform, however, may not be the optimal solution for any of them. In an effort to achieve a balance between highly-optimized solutions and reusable infrastructure, an overview of use cases in other disciplines will be developed. Understanding the methods and motivations for data access in such fields as Astrophysics (LSST, SDSS, Fermi) and Bioinformatics, and how this compares to HEP, should lead to an understanding of how far the underlying infrastructure can be generalized without compromising performance, and which aspects of the curation architecture need to be refined for a specific task. This particular task will benefit greatly from the inclusion of experts from the OSG as contacts for this proposal, since the OSG is the major US platform for multi-disciplinary computing in the US.

Survey of Technical Solutions for Archive Infrastructure. As a parallel activity, an effort will be made to survey and evaluate various possible components of a data and software archival system. This may include joint workshops and technical projects to describe, evaluate, and prototype some of the infrastructure needed for shared data archives, storage evolution, software packaging, distribution, and deployment, and data management.

As a clearer sense of the needs of the project emerges, focus will shift toward more technical aspects of enabling the Curation Challenge. Guided by the use case requirements, and in conjunction with the wider HEP community and ILDAP, preliminary versions of data description vocabularies and dictionaries, meta-data formats, and visualization techniques can be developed that will allow the efficient retrieval of physics data and software from an archive. Requirements and standardization for data, software, and analysis documentation, focused on this task, will also be required. In parallel, a hardware and software infrastructure can be designed that meets the criteria for long-term storage, interoperability with the European effort and OSG resources, and the particular constraints of the data-analysis challenge.

A final goal of the proposal would be to conduct the Curation Challenge well before the end of the funding cycle so that the results can be evaluated, disseminated, and published. As mentioned above, this effort will contain all of the ingredients of data and software preservation and access, but on a much smaller scale than that needed for the LHC experiments and other disciplines. As such, the problems encountered and solved can serve as a guide for the larger efforts that are necessary.

Outcomes

In the planning and prototyping activities the main outcome will be documents chronicling the decision path, technical design, and results from the Curation Challenge. Additional documents detailing use cases for HEP data and their implications for data and software reuse, access and preservation, the assessment of complementarity across multiple scientific domains, and the survey of technical solutions will also be a result of this work.

Broader Impacts

While it is unlikely that all disciplines can share a uniform common data curation/access infrastructure, several beneficial outcomes are possible from this effort. The identification of some commonality between the disciplines at all levels of the archive and access process can lead to the development of archive "modules" of storage, computing, and access infrastructure appropriate for whole classes of disciplines. With some effort, flexible metadata formats can be designed that would suit many different disciplines, allowing a common generic interface to extremely diverse datasets. The entire process of developing metadata for indexing, documentation, and retrieval, defining data use cases, and then arriving at an archiving solution can be codified for future efforts. The process of the "Curation Challenge" will thus become a template for other disciplines. This would enable any project in any discipline to know in advance "which questions to ask" when determining what archiving model is best for that particular application and which policy questions must be answered moving forward.

Activities

Overview: The initial effort will focus on a small number of workshops, presumably in conjunction with ILDAP, designed to solicit advice and draw upon the experience of the broader HEP and Multi-disciplinary communities to establish both a set of use cases and an understanding of the commonality and/or uniqueness of data curation solutions for HEP as contrasted with other disciplines. A parallel set of technical meetings will assess current best practice and the immediate future plans for data curation among HEP experts and multidisciplinary communities. A follow-up round of meetings will be necessary to reach agreement on prototype data/software description and metadata formats. As the technical path, guided by the derived use case(s) becomes clear, joint activities with the ILDAP effort will be staged to demonstrate interoperability of the US and European archive solutions. The Curation Challenge will be the culmination of both the technical and organizational research, combining a prototype hardware and software infrastructure with the higher-level data description, indexing, and access tools developed over the course of the program.

Work Plan for Year 1: Early in the project, a workshop including experts from other disciplines, possibly in conjunction with national or international meetings such as the American Society for Information Science and Technology's Research Data Access & Preservation Summit, the International Data Curation Conference, or the like. This meeting will address the questions of commonality in higher-level descriptions and access of data for various discipline-specific use cases. This would begin to address the question of what sorts of data need to be stored, and can a common metadata format be designed to allow generic access and indexing. An index of the policy issues that need to be solved for various implementations of data access and sharing would also be compiled. In parallel, a technical workshop will be held to evaluate the current and near-future techniques for various aspects of long-term data storage, distributed access, etc., also in conjunction with a national meeting of the appropriate expertise (e.g. Computing in High Energy Physics (CHEP), Association for Computing Machinery's Special Interest Group on Management Of Data (SIGMOD), IEEE Mass Storage Systems Technical Committee (MSST)). These would be followed with frequent video conference meetings whose purpose would be to crystallize the intellectual framework of the data preservation problem posed by the Curation Challenge:

- **Data Layer:** HEP data proceeds through many logical stages from raw detector output to presentable results. While the DPHEP data preservation model suggests four tiers of data, it is not clear that retaining access to and a description of Reconstructed Data, Simulation, and Analysis code (DPHEP Layer 3) is sufficient for reproducing an analysis. What data is really required?
- **Metadata and Query Model:** Best practices in long term data management require a separation of the logical structure of the data from its physical manifestation. To date, most HEP data processing activities have intermixed the two, which makes it difficult to move between technologies. What is the high level schema of the data, and what logical queries must be supported in the long term to enable easy access, visualization, and retrieval?
- **Software Specification:** For example, the current LHC software stack is well-managed but enormous and intertwined. To reproduce someone else's reconstruction currently requires access to the entire LHC software library, and a very large specification of how to configure the software and the operating system. How can we compactly describe what software must be used to reproduce a given result, even as operating systems an applications change? How does one archive, query, and retrieve a specification of the *computation* to be performed on the data?

To drive these initial discussions, we will conduct a Mini Reproduction Challenge in the first year. The challenge is simply stated: physics collaborators will select and process a small amount of data using standard HEP software, and then the computing personnel will be charged with manually reproducing that single result from scratch on a clean machine, using only the description given by the physics side. While far from a complete solution to the problem, it is a concrete deliverable that will force us to discover and understand many elements of the problem of reproducibility. The experience of the Mini Challenge will be documented and used to plan the following years of the project, including the applicability of this mode of inquiry to fields beyond HEP.

While this project will involve some degree of software/hardware prototyping and operation, we emphasize that the goal of the project is to outline the overall intellectual structure of the preservation problem, and leave the actual preservation activity to future efforts.

Personnel

As envisioned, this proposal is supported by the CMS and ATLAS offline, physics, and computing groups with the specific representation of expertise given by CMS offline and computing (Bloom, Hildreth), ATLAS offline and computing (Gardner, Neubauer). Specific computer science expertise is represented by Thain and Nabrzyski, who also heads the Notre Dame Center for Research Computing, which would host the hardware test bed. Specific liaisons with US ATLAS (Ernst at BNL) and US CMS (Bauerdick, Sexton-Kennedy at FNAL) computing and software efforts at the national labs have already been established to maintain close coordination with on-going activities and to tap the deep experience these Tier 1 facilities have in data handling and curation. In addition, a cohort of digital librarians with deep interdisciplinary experience is also represented (Long, Blair from Chicago, Johnson from Notre Dame) including specialists in Bioinformatics (Grossman from Chicago, Munn from Notre Dame). Coordination with the OSG is provided by two members of the OSG steering committee (Bauerdick, Ernst).

Budget

A preliminary budget is presented in the table below. Funds are included for two workshops each year, assuming $15k per workshop to defray costs and a small travel allowance for participants. Travel funds are included for a small number of people to attend international or national data preservation conferences. Personnel include a computational scientist and programmer from the Notre Dame Center for Research Computing (2 years, 0.5 FTE each), two Notre Dame graduate students, a Nebraska graduate student, a Chicago graduate student (all 0.5 FTE Year 1, 1 FTE Year 2), and a developer at Chicago (0.5 FTE, Year 2). All costs listed below include F&A returns to the universities. The Year 2 effort approaches 5 FTE. Note that we expect the need for computational scientists and programmers to ramp up towards the end of the first year, once the project has reached the point where technical specifications can be made. Costs are in $k.

	Year 1	Year 2	Total
Computational Scientist/programmer	84	154	238
Graduate Students	60	119	179
Workshops (2/year)	30	30	60
International Travel	15	15	30
Domestic Travel	10	10	20
Prototype Hardware	15	15	30
Total	214	343	557

International Support for this effort

Ian Bird, CERN, Director of the LHC Global Computing Grid

Jamie Shiers, CERN, Coordinator, ILDAP project; Director, CERN Grid Support Group

Steven Newhouse, Director, EGI.eu, European Grid Infrastructure

Cristinel Diaconu, CPPM/DESY, Chair, DPHEP

Ghita Radal, IN2P3, Director of Experiment Support

Fabiola Gianotti, CERN, Spokesperson, ATLAS Experiment

Joe Incandela, UCSB, Spokesperson, CMS Experiment

Pierluigi Campana, INFN Frascati, Spokesperson, LHCb Experiment

Paolo Giubellino, INFN Torino, Spokesperson, ALICE Experiment

Participants

Charles Blair, Director, Digital Library Development Center, University of Chicago
Kenneth Bloom, Department of Physics & Astronomy, University of Nebraska-Lincoln
Michael Fray, IT, University of Chicago
Robert Gardner, Computation Institute/Enrico Fermi Institute, University of Chicago
Robert Grossman, Institute for Genomics & Systems Biology, University of Chicago
Michael Hildreth, Department of Physics, University of Notre Dame
Rick Johnson, University Libraries, University of Notre Dame
Elisabeth Long, Associate University Librarian for Digital Services, University of Chicago
Natalie Munn, Department of Biology, University of Notre Dame
Jarek Nabrzyski, Director, Center for Research Computing, University of Notre Dame
Mark Neuebauer, Physics Department, University of Illinois, Champagne-Urbana
Chris Sweet, Center for Research Computing, University of Notre Dame
Douglas Thain, Department of Computer Science and Engineering, University of Notre Dame
Chuck Vardeman, Center for Research Computing, University of Notre Dame
Gordon Watts, Department of Physics & Astronomy, University of Washington

National Laboratory Contacts:

Lothar Bauerdick, LHC Physics Center, FNAL
Amber Boehnlein, Director of Scientific Computing, SLAC
Michael Ernst, US ATLAS Computing Coordinator, OSG Council, BNL
Liz Sexton-Kennedy, FNAL, DPHEP member, CMS Offline Coordinator