Data and Software Preservation for Open Science (DASPOS)

Michael D. Hildreth [#1] for the DASPOS Collaboration

[#] *Physics Department, University of Notre Dame*
Notre Dame, IN, USA
[1] hildreth.2@nd.edu

Abstract— **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. Recently funded by the National Science Foundation, the project is organizing multiple workshops aimed at understanding use cases for data, software, and knowledge preservation in High Energy Physics and other scientific disciplines, including BioInformatics and Astrophysics. The goal of this project is the technical development and specification of an architecture for curating HEP data and software to the point where the repetition of a physics analysis using only the archived data, software, and analysis description is possible. The novelty of this effort is this holistic approach, where not only data but also software and frameworks necessary to use the data are part of the preservation effort, making it true "physics preservation" rather than merely data preservation.**

I. Introduction

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.

The project described here, Data and Software Preservation for Open Science (DASPOS), represents a first attempt to establish a formal collaboration tying together US physicists from the CMS and ATLAS experiments at the LHC and the Tevatron experiments, scientists from other disciplines, and experts in digital preservation, heterogeneous high-throughput storage systems, large-scale computing systems, and grid access and infrastructure. The DASPOS activities will be connected into the DPHEP coordination effort, the experimental collaborations, and other related multi-disciplinary projects in Europe, Asia, and the US. Together, this group would represent the US in international efforts, acting as a coordinating point of contact, a partner in dialogue, and a technological consort. The intent of the project is to delineate and execute a set of initial, well-defined, small-scale activities to provide beneficial outcomes upon which a much more extensive, longer-term program can be based.

DASPOS represents an initial attempt to develop and specify an architecture for curating HEP data and software to the point where the repetition of a physics analysis using only the archived data, software, and analysis description is possible. The project was initially funded at the start of FY13 for three years. The novelty of this effort is this holistic approach, where not only data but also software and frameworks necessary to use the data are part of the preservation effort, making it true "physics preservation" or "knowledge preservation" rather than merely data preservation. This effort is an initial exploration of the problems to be solved at the technical, sociological, and policy levels, in order for integrated data preservation as envisioned by DPHEP to be possible. While it may not arrive at the optimal solution it will provide the solid foundation necessary for the next steps of preservation infrastructure development. The research involved is a combination of two overlapping activities: a "horizontal" coordination and consensus-forming activity, both internal to HEP and including other disciplines, to agree on prototype metadata definitions and other common aspects of data preservation, and the more technical construction of the "vertical" slice of archival infrastructure. One end result, the so-called "Curation Challenge" will be a small-scale but full system test of a particular archiving solution enabling the discovery and enumeration of the critical issues in establishing preservation architectures.

In addition to a technological demonstration, a primary result of this project will be a catalogue of issues that any effort in the preservation of Big Data will confront. A key aspect of this work will be the inclusion of different scientific disciplines in the discussions of research use cases, archival strategies, metadata definitions, and policy considerations. Through this extended dialogue, we will be able to establish elements of commonality that can lead to common technical and architectural solutions across disciplines. Equally important will be to discern at what level the disciplines diverge and at which point custom solutions to the archival problem will be required. We also expect to outline branch points throughout the preservation architecture specification where policy choices will dictate technical outcomes, leading to a blueprint for any discipline approaching the problems of

978-1-4799-3481-2/14/$31.00 © 2014 IEEE

large-scale data preservation and open access. We aim for these common solutions and principles established to serve a similar role within the HEP community that the OAIS model plays for Trusted Digital Repositories. Of equal importance to these broad-ranging policy and technology issues will be the training of a team of graduate students in the technical aspects of large data set preservation, global grid-based access tools, and other facets of this multi-disciplinary problem. Finally, the development of technologies for the preservation of large scientific data archives opens up the possibility of future scientific opportunities and insights not otherwise available.

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 HEP experiments and our colleagues from other disciplines. These include the following considerations:

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 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, 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.

II. 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.

III. 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.

IV. ACTIVITIES

Overview: The initial "horizontal" effort is focused on a small number of workshops 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 European efforts 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.

A. Workshops:

The project will hold three workshops annually, the first two of which were already held in 2013.

Workshop 1: Establishment of Use Cases for Archived Data and Software in HEP

Attendees: Participants from all of the HEP experiments considering long-term data preservation and access issues and digital librarians from participating institutions

Organizers: A team consisting of the digital librarians from University of Chicago and Notre Dame and HEP physicists from Notre Dame, University of Chicago and University of Illinois at Urbana-Champaign will take responsibility for facilitating the workshops and producing the outcome reports.

Location: CERN

Purpose: (i) Establish use cases for data access and re-use, especially for the larger DPHEP data tiers, since this will be a primary driver of the preservation architecture, (ii) define what data and associated information supports the use cases, and (iii) identify a preliminary set of metadata that would serve the needs of the HEP community in accessing the various forms of archived data/algorithms.

Workshop 2: Survey of Commonality with other Disciplines

Attendees: Broad participation from many NSF supported science efforts.

Location: Satellite workshop at IEEE/JCDL, Indianapolis, IN, July 2013.

Purpose: (i) Explore areas of commonality and difference, (ii) identify common metadata standards that could be designed to allow generic access and indexing of cross-disciplinary research data, and (iii) identify cross-disciplinary services that would support data preservation (e.g. software repositories).

Inputs: A discussion framework similar to that developed for the HEP-focused workshop will be developed for the cross-disciplinary workshop and will also be used to conduct individual or small group discussions with targeted colleagues who might not be available for the workshop (e.g. the researchers and technical staff involved in archiving the Sloan Digital Sky Survey-II).

Future workshops will embrace such core issues as Data Models and Query Semantics, Software Sustainability, the impact of Preservation Policies, and Storage Architectures. Technical reports summarising workshop findings can be found on the DASPOS web site: www.daspos.org.

B. Technical Development:

In parallel to the information-gathering and synthesis efforts, the technical side of the project will explore two aspects of the knowledge preservation problem from a perspective based on computer science and digital curation.

1) Data Model and Query Semantics: In order for a body of data to remain useful over the long term, the user community must have a common understanding of the organization of the data (the data model) and the form of queries that can be performed on that data model (the query semantics). To this end, a primary objective of this project will be to clearly document the data model and query semantics of the LHC data, gain the acceptance of the stakeholders for the model, and verify its ability to represent multiple physical organizations and software technologies. Each of the various LHC experiments has addressed this problem in some detail for its own purposes, however it remains an open problem to describe these sources in a common way that can be integrated into both local and institutional repositories. An early goal of this effort is to demonstrate a working prototype of the data model and query language on a fixed data set stored on a local filesystem independent of the LHC infrastructure, which will be followed by an expansion to demonstrate functionality with distributed storage.

2) Preserving and Evolving Processing Environments: Selecting data for preservation is only part of the challenge. At the LHC, like experiments in other domains, data must undergo reconstruction, analysis, and processing by a number of software tools that are under continuous development and refinement. A rich ecosystem of software and services are available, ranging from low-level system libraries for performing I/O, numerical libraries for physics reconstruction, and graphical systems for visualizing the results. Like the actual data, software and other artifacts must also be preserved, but the goals and the mechanisms of preservation vary substantially. Maintaining usability of data for processing requires sustained preservation and operation of a number of components comprising the required execution environment: user software, and dependencies; shared (experiment-wide) software, and external package dependencies; execution platforms (operating systems, compilers, and other utilities); data artifacts (of various types, in various formats); and associated services for computation. We will explore the solution space of program semantics by implementing multiple methods of creating and preserving a software execution environment, whether by deploying virtual machines, constructing software from source on demand, or some combination of the two.

In order to gauge the success of this project, we plan to create an Auditing Team that will define measurable tests of the performance of the various pieces of infrastructure that will be created.

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.

C. *Curation Challenge:*

For the first two years of the project, the work on the data model and the program semantics can proceed relatively independently. In the final year, we will bring the two pieces together to complete the *Curation Challenge*. At a high level, the challenge consists simply of allowing one team to completely reproduce the physics results of another, relying only on the high-level logical description of both the event data and the software to be applied. The Curation Challenge is shown schematically in Figure 1. After gaining some experience with the data model, the query language, and the high level software and processing specification, the audit group will develop a handful of challenge tasks that specify a data set via the query model and an analysis task via the program description. The technical group will carry out each of the curation challenge tasks and return the results to the audit group to verify the correctness. The final report will detail the nature of the curation tasks, the success or failure of individual curation tasks, the human and computational resources necessary to complete each curation task, and any missing pieces prohibiting successful completion of the Curation Challenge. The Curation Challenge has three primary components:

The Curation Process

This is the demonstrator process whereby the semantics and complete specification of all data artifacts, software, processing environment, validation recipes and target user analysis tasks are captured and stored in the prototype service. The *Auditing team* is to take a primary role here, working with the experiment processing and analysis experts to capture realistic processing steps and analysis results. The semantics developed as part of the DASPOS reference architecture are used to describe a particular curation challenge task, and to communicate this task to the *Technical execution team*. An important ingredient in the curation process is the capture and expression of a pre-defined, small scale set of validation tasks that can be run periodically over putative input datasets in automated fashion in a full-scale production system. Reference outputs (plots, histograms, numbers) will be stored and can be compared to validation runs later during the challenge.

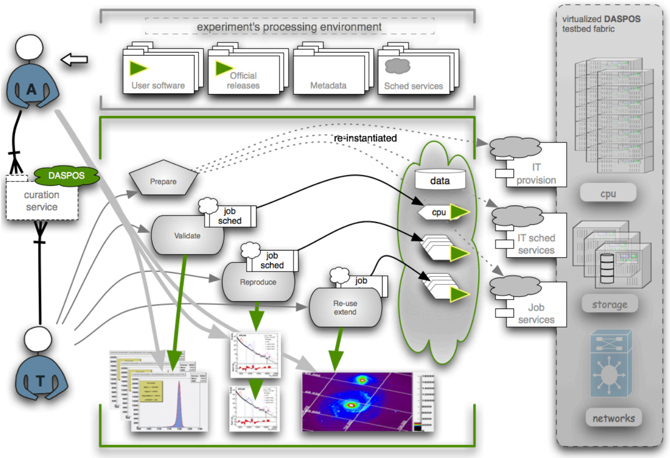

Figure 1: *Curation challenge* overview: The auditing team (A) prepares the prototype DASPOS curation service with necessary semantic description of the tasks, software and processing environment; a technical execution team (T) uses these data to marshal the processing environment from the DASPOS testbed fabric, and performs successively a validation, reproduction, and an extended analysis which is later audited for accuracy by the A-team.

Validation and Reproducibility Processes

The Technical team receives a description of the curation analysis challenge in the context of DASPOS semantics and reference architecture. A back-end virtualized testbed will be created to support this activity, providing the Technical team the ability to re-instantiate the processing environment according to the curated specification. Having the prerequisites in place the team can launch validation tests according to the specification. Both Technical and Auditing teams can check for validity. If initial validation tests are successful, then larger scale analysis reproduction is attempted, again launched and managed by the Technical team. This could involve more complex analysis over larger scales of preserved data. Bit-by-bit (approaching checksum comparisons) or higher level confidence level tests over sets of distributions will be used to determine if reproducibility criteria are met by both Technical and Auditing teams.

Re-use, Extend Challenge Processes

In the re-use/extension case, the Technical team re-uses preserved input datasets extending the original analysis and perhaps performing an analysis with significantly different algorithms. In this case the results of the Technical team's executions will be independently analyzed by the Auditing team for accuracy of computation.

V. Personnel

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). A variety of data and analysis preservation activities, as well as CDF and ATLAS, are represented by Cranmer. 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. The DØ data preservation efforts are represented by Watts. Specific liaisons with US ATLAS (Ernst at BNL) and US CMS (Bauerrdick, 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). The list of participants follows:

Kenneth Bloom, Department of Physics & Astronomy, University of Nebraska-Lincoln, Kyle Cranmer, New York University, 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, Head of Digital Library Services, University of Notre Dame, Elisabeth Long, Associate University Librarian for Digital Services, University of Chicago, Natalie Munn, Center for Digital Scholarship, University of Notre Dame, Jarek Nabrzyski, Director, Center for Research Computing, University of Notre Dame, Mark Neubauer, Physics Department, University of Illinois, Champagne-Urbana, 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, Lothar Bauerdick, LHC Physics Center, FNAL, Michael Ernst, US ATLAS Computing Coordinator, OSG Council, BNL, Liz Sexton-Kennedy, FNAL, DPHEP member, CMS Offline Coordinator.

Acknowledgment

This effort is generously supported by the National Science Foundation through award PHY-1247316.

References

[1] The DPHEP Study Group, "Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics", DPHEP-2012-01, May 2012, arXiv:1205.4667v1.