Data and Software Preservation for Open Science (DASPOS)

Report for Workshop 1:

Establishment of Use Cases for Archived Data and Software in HEP

March 21-22, 2013

CERN, Geneva, Switzerland

M. D. Hildreth, E. Long, R. Johnson, DASPOS co-organizers, with

K. Bloom, R. Gardner, M. Neubauer, D. Thain

Abstract:

The first DASPOS meeting was held joint with the 7th DPHEP Workshop, with planning coordinated between DASPOS and DPHEP management. It was hosted by CERN. Most of the first day of the workshop was devoted to areas of mutual DPHEP/DASPOS interest. This included projects related to analysis preservation, such as presentations on data-analysis-based outreach activities, Rivet/HEPDB, and an analysis preservation effort led by phenomenologists. It also included overviews of current data/analysis preservation efforts from Babar and the Tevatron experiments, and an overview of data-analysis workflows from the four LHC experiments. This report summarizes findings and areas of future work based on the information presented and discussions during the workshop.

This work was supported by grant NSF-PHY-1247316

(Updated Spring 2014)

1. Introduction

The first DASPOS workshop was held in conjunction with the 7th DPHEP workshop at CERN on March 21-22, 2013. Topics of discussion and the talks themselves were arranged thematically such that those of strong mutual interest to DASPOS and DPHEP were grouped together on the first day of the workshop. Participants included representatives of each of the four LHC experiments, the Tevatron experiments, BaBar, and DESY. In addition, a separate set of discussions on outreach and high-level analysis preservation also included representatives from the theory community and Rivet/HepData.

The workshop agenda, including copies of the talks that were presented, can be found on the public link: http://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=233119

1.1 Workshop Themes

Part of the scope of the DASPOS project is to explore various aspects and levels of data and knowledge preservation in High Energy Physics and beyond. A primary focus of this first workshop was an examination of the data processing and analysis workflows of the active HEP experiments. This was coupled with a review of other "high-level" analysis preservation efforts and projects related to outreach. These other activities were included because they might serve as viable methods for preserving analysis knowledge, functionality, and documentation.

More specifically, the sessions included presentation and/or consideration of the following:

- Outreach efforts, data formats, visualization tools, presented by the four LHC experiments
    - What has been the progress of each experiment in developing exercises, what data is used, what technologies are involved?
    - Is there any interest or benefit to common tools or formats for analysis exercises and for data visualization, both for histograms and event displays
- Analysis Preservation embedded in common HEP Tools
    - Rivet, HepData
        - Uses, content, potential expansion to be more inclusive
    - Les Houches Recommendations for presentation/dissemination of analysis results for new particle searches
        - Efforts at standardization, preparation of analysis database
    - RECAST
        - Not explicitly presented at the workshop but will be discussed in the context of high-level analysis preservation below
- Data Processing and Analysis Workflows of HEP Experiments
    - What are the common elements and differences in the way each experiment
        - Sets up the processing environment for its data, including conditions
        - Handles the data processing itself, including processing steps, workflow creation and storage
        - Analyzes the data, including additional processing steps and the use of common formats
        - Preserves knowledge about finished analyses

The overview of the workflows and analysis efforts in the experiments was facilitated by the development and distribution of a Data Preservation questionnaire based on the Data Curation Toolkit (ref?). The full questionnaire is included as an appendix to this report.

1.2 Workshop Goals

As stated in the original DASPOS proposal, the first workshop was intended to:

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. (ii) define what data and associated information supports the use cases, and
iii. (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.

To this end, a large fraction of the workshop was devoted to detailed presentations from each of the experiments on processing and analysis workflows. Analysis of the inputs will be presented in the later sections of this report.

1.3 Overview of Report

The report follows the general flow of the themes outlined above. While this report does contain some summary of the factual information presented in each of the sessions, its main purpose is to distill the wide variety of information into a form that can be understood by specialists from other fields besides High Energy Physics. To that end, each section gives a short summary of the content presented in the workshop and then elaborates on the overall themes that either arose from the discussion or that have developed from further reflection. Section 3 examines the different processing and analysis workflows presented by the experiments. An overall analysis of knowledge preservation in HEP is presented in Section 4.

2. Uses of Level 2 Data

An entire session of the workshop was dedicated to the exploration of the uses and access of Level 2 data, which, in the DPHEP nomenclature, refers to actual data and simulation presented in higher-level simplified formats. Many experiments have outreach programs based around analysis of a subset of their data using custom-developed tools, such as analysis portals and/or event displays. While not explicitly containing data, tools such as RIVET and RECAST encapsulate the actual algorithmic content of an analysis, making analyses reusable for future studies or comparisons. By exploring the various different efforts making use of Level 2 data, it was hoped that some common themes might emerge, motivating a possibility for common infrastructure. A second theme was that of high-level analysis preservation: to what extent are these frameworks, either for outreach or for broader community use, suitable as a means of preserving the details of a physics analysis? An overview of the topics presented and a discussion of these themes are presented in the sections below.

2.1 Level 2 Data for Outreach

All of the LHC experiments have developed outreach efforts based on giving members of the public access to small samples of data, and perhaps accompanying simulated data. These efforts are designed to give those outside of HEP some taste of what a particle physics analysis is and how it is conducted. To this end, they necessarily contain extensive documentation about

how to do a small set of analyses, albeit in a simplified environment. They also have a well-documented means of transforming the full data format(s) of a given experiment into a simplified format suitable for these applications, as well as an easily-understandable description of the contents of the format itself. Due to the limited resources typically allocated to these efforts, however, each experiment has understandably followed their own paths of least resistance toward establishing these outreach activities. This has resulted in many different incarnations of data formats and infrastructure, as can be seen in Table 1, below, which summarizes the salient features of the outreach efforts. (Updated in 2014)

	Alice	**Atlas**	**CMS**	**LHCb**
Event Display(s)	Root-based, 2nd simplified one?	Java-based	iSpy (http://cern/ch/ispy), browser based using jQuery & pre3D, ruby plugin in SketchUp, C# API for Unity	OpenInventor
name		ATLANTIS, VP1	iSpy	Panoramix
format of Geometry description	(1) Root, 2nd simplified one?	XML, full Geometry	XML/JSON	XML
Data Browser /Histogrammer/ Demonstration analyses	X/Root-based – looks like LHCb one. thinking of browser one w/o Root.	MINERVA, HYPATIA, LPPP, CAMELIA, OPIoT	Java-script based tools	X-based
Data Format(s)	Root	Jive-XML, Root, Full EDM, AOD, xAOD	ig	Root
self-documenting?	?	XML one is	Y (http://cern.ch/ispy/ig-specs.htm)	?
Master Class uses	various very specific analyses, some based on V^0s, others on general tracks	W, Z, Higgs, including large MC samples and data	similar to ATLAS, different datasets, not so much MC	D lifetime
Comments:	Root too heavy for classroom use			

Table 1. An overview of the different features of the outreach efforts from the four LHC experiments (only the LHC experiments were consulted for outreach details.) See text for a description of the categories listed.

Given the organic way the outreach technologies have been assembled, it would likely take substantial effort to move toward a base common format. Alice and LHCb might, however,

benefit from the development of a more general outreach architecture, especially for event displays.

A more general outreach architecture, perhaps based on a common format, common event display, and a "converter" that would allow access to multiple experimental datasets could have broad appeal. It would allow the easy comparison of data from different experiments on a common platform, with common tools, without the investment of having to learn several different event display or analysis package software suites. Such a converter is being developed for the outreach project in Finland that will use the CMS public data release. Here, a thin layer of software will convert data in a relatively low-level format (called AOD, see section 3) from the CMS experiment into a simplified representation that can be used for further analysis or visualization using an event display that consumes this simplified format.

2.2 Outreach Efforts as a Mode of Analysis Preservation

The tutorials and "master classes" that use outreach datasets are perhaps the most completely documented analyses in the high energy physics domain. In addition, each experiment also maintains a set of internal tutorials designed to introduce new members of the collaborations to the intricacies of the experimental software and data analysis. In all of these cases, the ancillary documentation, data files, and associated software allow a target population (outreach students, or, in the case of training tutorials, knowledgeable students or postdocs) to replicate analyses of varying levels of sophistication. In the sense that replication is possible, these analyses have been "preserved". However the means of "preservation" varies, from transient web or Wiki pages to printed materials. None of these modes of preservation would fit the characterization of proper curation of a preserved analysis. These analyses may be able to serve as examples of the range of descriptive content, processing code, and workflows that are necessary ingredients for analysis preservation. Certainly, as knowledge preservation infrastructure is being developed, these analyses can act as test cases for different representations or abstractions of the analysis process.

2.3 Level 2 Data for Re-Analysis

Several efforts exist that allow the "preservation" and repetition of an analysis at a relatively abstract level. The most established of these is the **RIVET** framework (https://rivet.hepforge.org/, RIVET stands for Robust Independent Validation of Experiment and Theory.) RIVET currently allows the comparison between experimental observables in particle physics and the theoretical predictions produced by theoretical models incorporated into various Monte Carlo generator codes. The framework is valid as long as the measurements have been corrected for the smearing introduced by detector resolution effects, noise, reconstruction efficiencies, etc., so that they can be compared directly with the theoretical predictions. Any Monte Carlo output can be juxtaposed with the data, as long as it can produce output in HepMC format (footnote: ?). A series of standard tools written in C++ can be exploited to replicate analysis cuts and procedures within the RIVET framework. Once validated, the analysis "code" can be included in the RIVET distribution, allowing anyone to reproduce the results of the analysis using independent Monte Carlo generation. RIVET is distributed as a software package with accompanying data from the included analyses. Note that the RIVET infrastructure, while relying on standard formats for the exchange of data, has essentially been written by those in the RIVET project; including an analysis in the RIVET repository requires a translation of the analysis software used to produce the results into the RIVET standard.

A second means of analysis preservation is represented by the **HepData** archive (http://hepdata.cedar.ac.uk) hosted by the Institute for Particle Physics Phenomenology (IPPP) at the University of Durham, UK. Its main repository is the "Reactions Database", which contains results from HEP experiments. The type of result can vary from total and differential cross section measurements to acceptance/efficiency grids in mass parameter spaces for Supersymmetry searches. It represents another level of detail beyond what is typically available in INSPIRE (http://inspirehep.net/), for example, but it does not usually preserve the code necessary to reproduce the analysis. As an aside, INSPIRE entries often contain links to entries and additional information in the HepData archive.

A third example is the **RECAST** framework (K. Cranmer, I. Yavin, JHEP 1104:038,2011.) More elaborate than the HepData and RIVET efforts, RECAST incorporates a full experiment analysis framework and the capability to generate events from new physics models, then subject them to a simulation of the particle detector and its reconstruction algorithms. The processed events can be analyzed using the same algorithms that produced a published result, and the results can be compared with those from collision data to constrain the new models in question. The RECAST structure includes a "front end" interface to the outside world where those interested in re-using an analysis can submit requests and inputs used in the processing. The RECAST API would mediate between the user interface and various capabilities provided by the "back end" processing installation. The back end does all of the processing and analysis work, and the results, if approved, are returned to the user.

At the workshop, an important use case for data/analysis preservation was made in a presentation by S. Sekmen describing the Les Houches recommendations for the presentation of LHC results. (Eur. Phys. J. C 72 (2012) 1976; arXiv:1203.2489). This document aims to create a standard way in which searches for new physics and the data on which they are based are presented in order to allow easy re-use by phenomenologists. The Les Houches publication contains many recommendations, but the ones most relevant to this current discussion concern analysis preservation and access to necessary data:

> "**Recommendation 1a**: Provide a clear, explicit description of the analysis in publications. In particular, the most crucial information such as basic object definitions and event selection should be clearly displayed in the publications, preferably in tabular form, and kinematic variables utilized should be unambiguously defined. Further information necessary to reproduce the analysis should be provided, as soon as it becomes available for release, on a suitable common platform.
>
> **Recommendation 1b**: The community should identify, develop and adopt a common platform to store analysis databases, collecting object definitions, cuts, and all other information, including well- encapsulated functions, necessary to reproduce or use the results of the analyses, and as required by other recommendations."

It was claimed that attempts are underway to both create an informal common analysis database for analyses developed by phenomenologists and to define a common code format for describing analysis algorithms.

Incidentally, an example was shown of an ATLAS search analysis with a very large amount of information uploaded to the HepData repository. Since HepData can accept data in many formats, it is not surprising that it can accommodate the sorts of information needed to replicate a new particle search, yet this use case lies quite far from its original intent as a repository of cross section measurements.

2.4 Discussion of Level 2 Analysis-Preservation Activities

The strongest use-cases presented for the high-level preservation of analyses come from those theorists wishing to re-run an analysis on a new model in order to understand what constraints existing data places on new physics ideas. Demand exists for a central repository of analysis algorithms and the information needed to reproduce analyses.

One potential solution to this need exists in the RECAST framework, which we will refer to here as a "closed" system. In principle, a single experiment would set up a RECAST installation that incorporates its analysis code and simulation framework. None of this code base would be exposed to the outside world, leaving the experiment in complete control of the content and function. The experiment would also have complete control over which analyses were allowed to become public based on comparisons with experimental data. Within the RECAST framework, standard formats (based on ?) for analysis algorithms, background distributions, and other elements needed to reproduce analyses are required in order to insure automatic inter-operability of different analyses within the framework itself. Thus, once all of the ingredients of an analysis have been put into RECAST, it is "preserved" in the sense that it can be re-run at any time using a standard (or new) set of inputs, either for validation or exploration of new ideas. While not the intended use, it would also be possible with some re-configuration to re-run the analysis on different or new data.

The RECAST framework presents several positive features when looked at from the perspective of analysis preservation:

- All ingredients of an "ingested" analysis are preserved in a common (to RECAST) format, along with instructions on how to run the analysis.
- The analysis can be re-run at any time. The outputs could be used, for example, for validation purposes.
- The level of detail at which an analysis can be reproduced is significantly enhanced by the capability to run the full detector simulation and reconstruction. Essentially, the full code base and executables from the experiment are encapsulated in the RECAST back end processing.
- Control over the use of the framework by outside entities rests entirely with the experiment. Because none of the analysis ingredients are accessible by the public, there is a natural mechanism by which a "gateway" to the physics might be realized. This may or may not be an advantage, depending on how "open" the interpretation of "open access" is taken.

Several aspects of the RECAST framework suffer from common potential problems of knowledge preservation:

- The resources of the computing back-end will need to evolve over time. Since whatever processing is provided is tasked to run a full suite of detector software, including simulation and reconstruction, the full experimental code base must be migrated to new computing platforms when such transitions become necessary. The entire set of processes must be kept functioning in order for the RECAST framework to produce appropriate results. If individual experiments are maintaining their own RECAST installations, significant effort will be required to maintain the functionality of the framework.
- Since each experiment would potentially maintain its own RECAST framework, the prospects for common solutions might be more limited. In principle, one could envision

a central RECAST repository with code bases from all participating experiments. In practice, this may be quite difficult to achieve.

- The "closed" code base and the overall structure limit access to the preserved knowledge to those who control the repository. Since they are the original experts from the experiments, this might be appropriate. Without different access methods, especially some sort of "back door" that members of the collaboration can use to access the analysis repository, the RECAST framework would be difficult to use as a resource for the experiment itself.

In contrast, the RIVET framework has been used as a *public* repository of analyses for many years. Analysis preservation consists of reproducing the analysis algorithms using the utility functions provided by the RIVET framework and uploading experimental data in a suitable format. Originally designed as a tool to compare various Monte Carlo generators with unfolded data distributions, it continues to be used primarily for this purpose. Once an analysis is put into RIVET, however, anyone can examine the analysis code and the reduced data provided for comparisons.

Compared with RECAST, however, the current functionality of RIVET is somewhat limited. Some of the differences are listed here:

- Because it was designed to study "background free" analyses related to the details of QCD, it is currently configured merely to provide simple comparisons between distributions obtained from a single generator source after a set of analysis cuts and manipulations.
- More advanced analysis capabilities, such as background subtraction, are not implemented.
- There is also no way to include a detector simulation, or even the degradations in resolution and particle collection efficiencies that the interaction with the detector will introduce.
- The level of interpretation that is possible does not include limit-setting, likelihood fitting, or other more advanced analysis or statistical techniques.

In the context of analysis preservation, however, RIVET could, with some further development, offer several advantages. Despite the issues raised in the previous paragraph, the current implementation of RIVET offers several positive attributes:

- RIVET is widely used throughout the High Energy Physics community as a means of sharing analysis results that could be useful for future understanding of QCD parameters. It represents a common standard for analysis preservation in this sense.
- The RIVET installation is quite "light" from a footprint standpoint. The code base is small and runs on essentially any platform
- The code base is open source and is based on widely-used HEP conventions. The tools provided are flexible, and new features can easily be added.

An interesting development that DASPOS will facilitate is to create a connection between RECAST and RIVET. It should be relatively straight forward to create a "back end" for RECAST such that any analysis implemented in RIVET could be subject to the RECAST framework. This could offer one avenue towards making the advanced tools of RECAST available to RIVET analyses. This interoperability is something to explore moving forward.

A second strong impetus for knowledge preservation comes from the experiments themselves. Whether in the context of preserving the ability to repeat an analysis for physics comparisons with a future dataset, for an archival record, or for validation purposes, the experimental community has an interest in establishing procedures and infrastructure for analysis preservation. The internal needs for the experiments fall much more closely to the "complete preservation" model, however, than towards a simplified Level 2 data approach like those discussed here.

What, then, appears to be the best way forward to satisfy the desires of the theory community for analysis re-use? The level of detail specified in the Les Houches publication is well beyond that of a typical Level 2 data description. Developing the infrastructure for an analysis database of this complexity with the flexibility to capture the necessary nuances of the experimental work is a non-trivial undertaking. It is possible, however, that the creation of such a repository would simplify generic analysis preservation, since it would operate at the abstract level of analysis objects, rather than the preservation of a specific code base.

3. HEP Experiment Workflow Analysis

While the Level 2 data preservation and outreach discussion is important, the main focus of the workshop was to arrive at an in-depth examination of the data-processing and analysis workflows of each of the HEP experiments. Additional information beyond the workshop presentations was provided by asking each of the experiments to prepare answers to a "Data Interview Template" created prior to the workshop and submitted to each experiment in advance. A copy of the template is provided in Appendix A to this report. The interview template provided a framework for the experiments to outline their thoughts or plans for data/software/knowledge preservation using a common set of considerations. To our knowledge, this was the first forum where each of the LHC experiments presented detailed analyses of their analysis workflows and their data preservation efforts.

3.1 Some HEP Terminology

The basic logical unit of data in particle physics is called an "event". The information in an "event" corresponds to all electronic detector signals originating in a single "interaction" in which, in the case of the LHC and other storage rings, two counter-rotating beams collide to produce new particles. The detector elements are typically sensitive to recording particle interactions for times from a few nanoseconds to times of order a hundred nanoseconds around the time of the collision. The shorter times are comparable to the propagation time for a particle to reach the outer edge of the detector, so the majority of the signals recorded come from particles from the primary interaction.

The probability that a given physical process occurs during a single interaction is given by the relative probabilities of the myriad final states that can be produced by the collisions of the beam particles. The individual collisions thus sample all possible final states at random. Since desired experimental signatures, such as a Higgs boson decay, rarely are unique, it is impossible to identify the products or progenitor of a given event with 100% certainty. Because of this, all high energy physics studies are statistical in nature, where ensembles of events are considered

and properties of the ensemble are measured. Discovery of new particles, for example, occurs when statistically-significant evidence for a new process, using whatever observables that have been defined, is seen above the level expected for other known processes. Because of the random nature of each event, the data from a single particle collision is of no use for physics analysis. Large samples of events must be compiled and filtered in order to produce sensible physics. Essentially all particle physics analyses are measurements of ensemble properties of these event samples, or a subset thereof. The nature of the science requires the reduction and processing of large datasets in order to extract small numbers of physics results, depending on the nature of the analysis being pursued. Because the particle detectors that record the collisions are designed to be generically capable of recording a broad range of interesting collisions, the variety of analyses that can be pursued using a given recorded dataset is quite large. So, many different analyses can be published by examining various aspects of a single set of collisions and applying different selection criteria or studying different observable variables.

3.2 Analysis of Workflows

Not surprisingly, the data processing and analysis workflows of the modern high energy physics experiments are remarkably similar. A generic outline of "typical" data processing is given here, with caveats and differences between experiments discussed below.

Each experiment begins with a "Reconstruction" step consisting of mainly the application of pattern-recognition and local-maximum-finding algorithms that convert the "raw" binary data read out from the detector elements into recognizable "objects" (i.e., particle trajectories, clusters of energy depositions in calorimeters, etc.). Further refinement of the interpretation of these objects is also done, resulting in the creation of "candidate physics objects" (electrons, muons, particle jets) that are combinations of the basic objects. The Reconstruction step thus forms the basis of the physics interpretation of a given particle collision. The data produced by the reconstruction step can contain a wealth of detail, from the original individual processed hits in individual detector channels all of the way through the various intermediate stages to the final "refined" physics objects. After the initial commissioning phase of an experiment, most of the basic and intermediate data categories are discarded, and only the refined objects necessary for further analysis are kept. This reduced output data is usually given a name identifying it as the basis for analysis, such as the Analysis Object Data or AOD.

Since the information contained by the AOD data tier is sufficient to serve as the basis for many physics analyses, some analyses use this data as the primary input to their final analysis routines. Often, however, one or more intermediate processing steps are desired in order to perform additional computations on the refined physics objects, to drop uninteresting events, and to discard extraneous information that may be irrelevant to a given physics analysis. Since the AOD data for a given event tends to be relatively large, both the dropping of events (known as "skimming") and the reduction of the event content (known as "slimming") result in a reduction of the final data size that will be useful for analysis. One or a series of slimming/skimming steps results in a final analysis data format that is usually customized to the needs of a particular individual or analysis group. From here, the final calculations can be performed that serve as direct input to the physics results contained in a publication.

An important consideration, to be explored, is the number of external resources that is required in order to perform the various processing steps. Generally, the Reconstruction step requires at

least one and sometimes many different databases that store all manner of calibration constants, conditions data, etc., that are required for interpreting the raw data of a given interaction with the highest possible accuracy. Once this initial processing is finished, however, dependencies on external databases or other sources of information become much weaker and potentially harder to quantify. Enumerating and potentially encapsulating these external dependencies will be an important ingredient in the analysis preservation process.

Essentially, all of the experiments have remarkably similar processing workflows for the larger processing steps (Raw to Reconstruction, Reconstruction to AOD, Monte Carlo Generation). The details can be seen in the slides posted to the Indico agenda for the workshop[1]. There are very minor differences in constants-handling (Alice, for example, has text files that can easily be shipped around with the data, while the other experiments make more extensive use of database access from processing.

The post-AOD workflows, directed at physics analysis, is where there is the most variety of approaches. CMS, for example, makes extensive use of common data formats for analysis groups, each of which is derived from a centrally-used AOD format. ATLAS is much less central. The other experiments are in between in terms of commonality. It is true, however, that each processing step between the final centrally-processed format and some reduced format can be reduced to a logical skimming/slimming description. The final steps to produce publication-quality plots and the final results are sufficiently varied that direct preservation (i.e., capturing an executable, or the entire source/script code) is likely the only way to insure that these final operations are preserved.

There are some other issues that arose during the discussions. One was that of provenance retention in the derived datasets. Depending on how the processing is done, the parentage and computing (producer) description of a given file may not be included. If this is the case, and the workflow is to be preserved, an external structure to capture that provenance chain will need to be created.

4. Current State of Data Preservation Policy Statements at LHC

CMS: Data policy and intent to release data to the public was approved in 2013.
LHCb: Data policy and intent to release data to the public was approved in 2013.
ALICE: under discussion (2014)
ATLAS: under discussion (2014)

5. Conclusions

Because the opportunities presented themselves, this workshop was able to address several different, but interrelated issues. In terms of addressing the high-level possibilities for analysis preservation: “Level 2” type data (e.g., simplified format events, or encapsulated analyses) are used in a variety of different ways, by a variety of platforms, with no common formats. This is

[1] http://indico.cern.ch/conferenceDisplay.py?ovw=True&confId=233119

understandable, given the time scales of the development of the infrastructures that use this information, and the limited amount of effort available for development and maintenance. Some installations have existed for decades; others are newly-created – all have very few people involved in sustaining these efforts. Analyses captured in outreach efforts, while allowing broad public access to analysis frameworks and data, have to each be created "by hand" with extensive documentation for them to be useful. This is a huge overhead and is unlikely to be adopted as a common method for analysis preservation. Some of these other projects, however, such as RIVET and RECAST, have the potential to capture arbitrary analyses. RIVET, in particular, has been used to parameterize well over a hundred different analyses in a generic framework that allows comparison of unfolded data distributions with different Monte Carlo event generators. The RIVET framework allows the specification of a wide variety of analysis steps, meaning that many different analyses could be preserved in this fashion. This, unfortunately, also requires a large effort from the analyst for preservation, but it offers the benefit of a universal format and easy portability. An expansion of the scope of RIVET (i.e., dropping the requirement that its products and input are only unfolded (= "truth") distributions) could represent one possible way of having a universal language for analysis preservation at a high level of description, without the need for preserving individual analyst's software. A DASPOS project to connect RECAST with the RIVET framework is underway. This will significantly broaden the capabilities of both systems.

On the workflow front, there were no real surprises to those familiar with the standard High Energy Physics computing models. Each of the major experiments presented very similar processing and analysis chains, differing only in minor ways by data handling strategies and how external constants are handled. What distinguishes the HEP workflows from those of many other disciplines is the nested levels of processing required to go from the raw data written by the detectors/instruments to the final physics analysis plots and results. Central processing using very large facilities is essential to produce the simulated data and to reduce the raw data to a "physics-readable" form. Even though there is a diversity of processing that occurs in producing physics results, each of the subsequent steps can be well-defined semantically. The main issues moving forward will be related to creating a preservation framework that is easy to use, so that its adoption can be straightforward.

Data/Software Interview Template

Name:
Dept/Center:

A. Overview of the Data

5.1 1. Type and Extent

A. Description of Data:

B. Approx number of files:
C. Avg file size:
D. File format(s):

5.2 2. Data Lifecycle

How many stages will the data go through? Does the size/number/format of the files change at each stage?
e.g.
- Collection stage: e.g., *1000 Raw files from the detector*
- Analysis stage 1: e.g., *2000 processed files*
- Analysis stage 2: e.g, *200 sample files*
- Publication stage: *Summary data tables, ancillary information*
- Preservation stage: *Raw files + processed files + codebook + summary data+metadata*

5.3 3. Tools (Hardware/Software)

A. What tools are used in generating/collecting/processing the data?

B. What tools are required to utilize or analyze the data? *(e.g. Externals like ROOT; project-specific code)*

C. Are these tools widely used in your field? Are they proprietary? Are there alternatives?

5.4 4. Software Lifecycle

5.5 For each of the stages of the data lifecycle identified in question 2, please identify the software package(s) required to access and analyze the data.

5.6 A. For each stage, indicate whether the software is "external" (i.e., ROOT, databases, GRID software) to the central experiment software, or included within it. For external services, please indicate which pieces of information or functionality that they provide.

B. Indicate, if possible, which version of the software is required.

B. Data Management

5.7 5. Storage, Backup, and Disaster Recovery

A. What are the primary ways you currently maintain your data (including storage media; software tools, etc).

B. Do you currently make back-up copies of your data?

C. What, if any, security measures to protect your data?

D. Is a disaster recovery plan in place or procedures to avoid loss from technological failure?

E. Does your funding agency require a data management plan?

F. Data Management and Disaster Recovery Maturity Rating

1	2	3	4	5
Data management activities focus on the day-to-day	Some awareness of potential risks but few take preventative action	Policies and plans are in place for disaster recovery and long-term sustainability	Disaster recovery plans are accompanied by procedures for implementation Data loss, a break in the research process, or loss of access to data is unlikely	Disaster recovery plans are routinely tested and shown to be effective Succession plans (e.g. an alternative data centre) are in place to safeguard data

5.8 6. Data Organization/Description

A. How is the data organized? *(e.g. database tables, file hierarchy, etc.)*
How is that organization documented? *(e.g. a codebook; a data dictionary; column headings in a spreadsheet; etc.)*

B. Are there any standard data formats in your field or experiment? If so, are you using them for each step in the data lifecycle ?

C. Is this amount of organization/description sufficient for another person to use or understand your data from inside your experiment? From outside your experiment?

D. Data Description Maturity Rating

1	2	3	4	5
Metadata is an unfamiliar concept Low engagement with the need to document data	Metadata and data description practices vary by individual	Metadata is well understood and guidance is provided to support the use of standards Data are documented	Data are well labeled, annotated and systematically organized Data can be understood by other researchers	Metadata is routinely created and well managed The exemplary practice advances community standard

5.9 7. Software Organization/Description

A. How is your software organized? *(e.g. code repositories, work packages, etc.)*
How is that organization documented?

B. Is your entire software or a subset of it "versioned" in a controlled manner*? (e.g., "production releases" of Reconstruction code, online code, etc.)*
How is that versioning documented?

C. How are the versions of software relevant for each step in the data lifecycle specified?

D. Is this amount of organization/description sufficient for another person to use or understand your software from inside your experiment? From outside your experiment?

5.10

5.11 8. Data/Software Curation/Preservation

A. What are the most important parts of your data to preserve? *(e.g. raw data files; derived data files; etc.)*

B. How long would your data be useful or have value for you or others? What uses might people make of your data? *(e.g. other researchers could do different kinds of studies based on my data; replicability of my experiment for scholarly integrity; data would be of interest to historians of my field; etc.)*

C. In order for the data to be useful to future users, what software would have to be preserved?

D. If applicable, is the process for generating raw or derivative data and analysis documented, preserved, and reproducible? This would possibly include preserving any software or inputs.

E. Preservation Maturity Rating

1	2	3	4	5

Low awareness of requirements to preserve data	Data may remain available but mostly due to chance, not active preservation practice	Preservation is understood and well-planned	High levels of awareness and engagement e.g. data are selected for preservation and repositories are in place.	Data are efficiently and effectively preserved The infrastructure in place is understood, functions well and is widely used

5.12 9. Data Access and Sharing

A. For each data lifecycle stage identified in section 2, do you need/want to share your data? If so, with whom?

e.g.

- No one

- Project collaborators (are they at host and/or other institutions?)

- Host academic community

- Others in the field (e.g. through a disciplinary repository)

- Whole world (i.e. publicly available on the web)

B. When would you be willing to share? *(e.g. always; 6 months after analysis; 1 year after publication, once project is complete.)*

C. Would you place any conditions on the use of your data? *(e.g. registration; signed waiver; acknowledgement requirement; etc.)*

Data Sharing Grid			
Research Stage	**Who**	**When**	**Any Conditions**

D. Goals for sharing data

Who would be interested in your data? How would they use it? Does your funding agency require it?

F. Sharing/Access Maturity Rating

1	2	3	4	5
Individuals store data and manage access requests	Guidance and services are provided for data access but are poorly used	A mix of systems is in place to meet different access needs (e.g. shared storage, laptops, portable storage, commercial services) Security questionable	Access is systematically controlled through user rights and strong passwords	Systems meet all user needs (e.g. remote access, sharing with external collaborators etc) and security is maintained
Low awareness of data sharing requirements	Ad hoc data sharing occurs (e.g. data provided on request)	Data sharing is supported - training is provided and the necessary infrastructure is in place	Data are shared as appropriate (i.e. where legally and ethically possible) Support and infrastructure for data sharing functions well and is broadly adopted	There is a culture of openness. Data sharing systems are recognized and copied by others