Data and Analysis Preservation in PHENIX: the current outlook

M.Potekhin
Nuclear and Particle Physics Software (NPPS) Group

Review of the Plans of the PHENIX Collaboration
to Complete Analysis of the PHENIX Dataset
BNL
December 12th, 2019

Data and Analysis Preservation

- We'll use the term DAP to refer to data **and** analysis preservation
  - Data which cannot be analyzed is effectively useless
- Mandate from PAC
- DAP is commonly described as a union of
  - bit preservation
  - software infrastructure and application code
  - analysis know-how
- ...which implies knowledge management in all areas listed above
- Observing growing interest in the community, in the DAP and relevant tools over the past few years
- The goal is to have a reproducible analysis capability over a long period of time
  - also the capability to perform a modified or new analysis within the same framework
  - ...the goal if to ensure the continuity of the experiment's capability to perform analysis
- We shall review these components in terms of their current status, issues and mitigation
  - Focus mainly on "Tier-4" DAP (raw data and tools) - other levels do exist cf. "open data"

Bit Preservation

- PHENIX data on tape: approx. 24PB
  - A decision was made to have one copy hosted by SDCC (RACF)
  - cf. ISO 16363 (Trustworthy Digital Repository)
  - Most criteria are met (e.g. hardware and media upgrade policies and procedures etc)
- Disk space allocation is approx. 1PB (GPFS)
  - Currently used for production and many analyses
  - Will be used to host components of reference data within the scope of data preservation, especially those that will need to be accessed by the DAP Web services
  - Reasonable expectation of long term support and/or migration

Software Infrastructure

- Version control - core code managed with CVS (git/GitHub used for some interim work)
  - However many parts of the analysis software code are not
- Current use of containers in production
- A well-maintained system of builds (credits: C.Pinkenburg)
  - Daily builds, special builds activated during updates of certain toolkits and packages (e.g. ROOT etc)
  - Robust tools like Insure++ and Coverity deployed to check the code
  - Weekly scan-builds to identify potential problems
- Important components:
  - PISA (simulation toolkit) - based on Geant 3
  - Fun4All (the PHENIX software framework)
  - ROOT
- Software documentation is fragmented and often not up-to-date; lack of functioning tutorials
  - Clearly an issue that will need mitigation
- Going forward, in the DAP context - challenge of build and validation with new compilers and OS

Knowledge Management (KM)

- **Knowledge management** is the key to DAP
- Involves data discovery, infrastructure knowledge and analysis know-how

Analysis know-how

- Analyses involve multiple steps, custom software and sometimes intermediate/custom data components
  - e.g. where did this “dead channel map” come from?
- Keeping a *detailed record* of the analysis procedures is a non-trivial extra burden on the researcher
  - Getting a quality paper published is the top priority
  - Currently handled via formalized requirements to the analysis notes (see below)
  - Ideally needs dedicated support
    - Curation and documentation of the analysis procedures
    - Some of this work has started
- Knowledge management in this area is a labor-intensive part of DAP

Analysis Notes

- **Analysis notes** are required from PWGs prior to a decision about a publication
- Each note encapsulates the know-how of a particular analysis (at least it's the goal)
  - ideally self-contained description of the workflow
  - cross-references to other analysis notes if necessary
- Kept in a well-organized archive
  - however there is no standard for the note (or archive) format across experiments
- A note must be complete in terms of the physics content and adhere to specific requirements
  - in particular describe the software and the "flowchart" of the analysis
  - may include code reviews (a big plus)
  - ...at the same time, enforcement may be relaxed at times

Analysis Notes in PHENIX

- PHENIX has a well-designed “note template” and policies regarding the note content - more on that below
- Archive: the PHENIX Web application for note revision and archival was created a while ago and has a few limitations:
  - in reality limited to a single file, typically PDF (with links to revisions)
  - search function is fairly basic
  - custom software based on PHP (long term maintainability/security issues)

An excerpt from the Analysis Note Template

...which is effectively a list of requirements, presented here as an example:

A flowchart of the major analysis steps with code locations

The flowchart is a birds-eye view of your analysis flow, with references to the basic working directories. “Code location” here means your working directories on RCF. Ideally these all should branch off a single base directory (which also makes backing up the “snapshot” – see below – in HPSS easier). Important: if you did some of the work locally (e.g. on your laptop), you have to migrate the codes and relevant files to RCF when you are finished, and make sure it works there, too. Examples of directories to be specified:

- Directory for Taxi code
- Directory to run Taxi
- Directory of the Taxi output
- Directory for Simulation code
- Directory for Simulation output
- Directory for analysis code and macros that analyze the Taxi output
- Directory for analysis code and macros that analyze the Simulation output
- Any other directory used during the analysis
- Directory for the final data file(s) and macros to produce physics plots.

PHENIX Analysis Notes Archive - the query page

PHENIX

Analysis/Technical Notes query form

Search Form

Download Analysis note template here. Use this template to write new analysis note for preliminary requests and final journal publiation
Use this form to search for technical and analysis notes.
Select first the desired note type either analysis notes or techinical notes. Searches may be made by author, title, note number, submission date (year) or by keyword. The default is "All" and it displays the entire list of the selected note type. Either first name or last name can be used for author. The search string for "Search by Submission Date (Year)" should be of the form "yyyy" (e.g. 2012).

Use AN Submission Form to add a new analysis note and TN Submission Form for a technical note.

Type of Note: Analysis Notes

Search by: Select from the list **Search String:**

Search by Author:
(Start typing author's name (last or first) until the desired name appears in the list and then select.)

Run Number: Select from the list

Collision Species/Energy: Select from the list

Physics Working Group: Select from the list

Analysis Type: Select from the list

Search Now Clear Form

PHENIX Analysis Notes - the archive contents

Number	Date	Title	Authors	Key Word	Links
an1425	2019-10-08	Run15 pAu identified pion and anti-proton spectra	Weizhuang Peng, Julia Velkovska	PLHF, p+Al_200GeV, Run-15, identified particles	an1425 draft an1282
an1424	2019-10-08	Run15 pAu identified pion and anti-proton spectra			an1424 draft
an1423	2019-10-08	Run15 pAu identified pion and anti-proton spectra			an1423 draft
an1422	2019-10-08	Run15 pAu identified pion and anti-proton spectra			an1422 draft
an1421	2019-10-08	Run15 pAu identified pion and anti-proton spectra			an1421 draft
an1420	2019-10-08	Jet Analysis in Run 15 p+p Collisions	John Lajoie, Milap Patel, Marzia Rosati, Jonathan Runchey	HHJ, p+p_200GeV, Run-15, jets	an1420 draft
an1419	2019-10-14	Neutral pion R_AA in p+Al, p+Au, d+Au and 3He+Au using combined Run-5, Run-8 and Run-15 \pp reference (PPG202)	Gabor David, Axel Drees, Norbert Novitzky	Heavy Ion, He3+Au_200GeV, p+Au_200GeV, p+Al_200GeV, d+Au_62GeV, d+Au_200GeV, Run-5, Run-8, Run-15, Run-14, single high pT particles, identified particles	an1419.01 draft an1152 an1269 an1270
an1418	2019-10-25	Model calculation of nuclear absorption in J/psi production at backward rapidity in PHENIX	Anthony Frawley	Heavy Ion, p+Al_200GeV, p+p_200GeV, d+Au_200GeV, Run-15, Run-14, quarkonia	an1418.02 draft
an1417	2019-10-22	Low pT Direct Photon Production in Au+Au Collisions at 200 GeV Beam Energy	Gabor David, Axel Drees, Roli Esha, Wenqing Fan, Norbert Novitzky	Photon, PLHF, Heavy Ion, Au+Au_200GeV, Run-14, direct photons	an1417.01 draft
an1414	2019-09-18	Template for PHENIX Analysis Notes	Yasuyuki Akiba, Gabor David		an1414 draft
an1413	2019-08-26	PHENIX Run14, Run15, Run16 PC2/PC3 track matching recalibration	Qiao Xu	PLHF, He3+Au_200GeV, Au+Au_14.6GeV, p+Au_200GeV, p+Al_200GeV, d+Au_62GeV, d+Au_39GeV, d+Au_20GeV, p+p_200GeV, d+Au_200GeV, Au+Au_200GeV, Run-15, Run-14, Run-16	an1413.01 draft
an1412	2019-08-02	K^* production in U+U at $\sqrt{s_{NN}}$ = 192 GeV in Run12	Alexander Berdnikov, Yaroslav Berdnikov, Vladislav Borisov, Dmitry Kotov, Daria Larionova, Iurii Mitrankov	HHJ, PLHF, U+U_193GeV, Run-12, single high pT particles	an1412 draft an065 an1010 an1374 an1401 an1402 an770 an911 an964 PPG148
an1411	2019-08-02	Protons production in Run12 Cu+Au at sqrt(s_NN) = 200 GeV	Alexander Berdnikov, Yaroslav Berdnikov, Dmitry Kotov, Mariia Larionova, Iurii Mitrankov	HHJ, PLHF, Cu+Au_200GeV, Run-12, identified particles	an1411 draft an1074 an1231 an1260 an1374 an683 an814 PPG146
an1410	2019-10-24	Direct, elliptic and triangular flow of π0 → γγ in d+Au collisions at 200 and 62 GeV	Veronica CanoaRoman, Gabor David, Abhay Deshpande, Jaehyeon Do, Axel Drees, Tom Hemmick, Carlos PerezLara	Hadron, Light, PLHF, d+Au_62GeV, d+Au_200GeV, Run-16, correlations, identified particles	an1410.02 draft an1367 an1406 an1407
an1409	2019-06-24	Run14 AuAu EMCal Geometry Tuning	Gabor David, Axel Drees, Wenqing Fan	PLHF, Au+Au_200GeV, Run-14, direct photons	an1409 draft
an1408	2019-06-20	Measurement of subevent cumulant flow in Run15 p+Au and Run 16 d+Au collisions	Ronald Belmont, Qiao Xu	PLHF, p+Au_200GeV, d+Au_62GeV, d+Au_39GeV, d+Au_20GeV, d+Au_200GeV, Run-15, Run-16, correlations	an1408.03 draft an1273 PPG206 PPG221
an1407	2019-05-30	PileUp Rejection Criteria based on BBC	Veronica CanoaRoman, Jaehyeon Do, Carlos PerezLara	Global, Heavy Ion, p+p_200GeV, d+Au_200GeV, Run-15, Run-16, correlations	an1407 draft an1304
an1406	2019-05-27	Q vector calibration	Veronica CanoaRoman, Jaehyeon Do, Carlos PerezLara	Heavy Ion, d+Au_200GeV, Run-16, correlations	an1406 draft
an1405	2019-10-20	Final Results on Double Helicity Asymmetries in Charged Pion Production in Longitudinally Polarized Proton-Proton Collisions at sqrt(s) = 510 GeV	Yuji Goto, Byungsik Hong, Ju Hwan Kang, Sook Hyun Lee, TaeBong Moon, Ralf Seidl, Inseok Yoon	Spin, p+p_510GeV, Run-13, single high pT particles, identified particles, A_LL	an1405.04 draft
an1404	2019-05-08	Measurement and analysis of three-pion HBT correlations for 0-30% Centrality in 200 GeV AuAu collisions	Mate Csanad, Bálint Kurgyis	PLHF, Au+Au_200GeV, Run-10, correlations	an1404.01 draft an1187 an1244 an1288 an920 PPG194
an1403	2019-08-15	J/ψ as a function of p_T in small systems with Yue Hang Leung's Correlated Background: Run15pp and Run15pAu, Run15pAl, Run14^{3}HeAu Centrality	Matthew Durham, Anthony Frawley, Sanghoon Lim, Krista Smith	HHJ, He3+Au_200GeV, p+Al_200GeV, p+p_200GeV, Run-8, Run-15, Run-14, J/psi, quarkonia, lepton pairs	an1403.06 draft an1306 an1354 an1369 an1391
an1402	2019-04-04	K^* production in Cu+Au at $\sqrt{s_{NN}}$ = 200 GeV in Run12	Alexander Berdnikov, Yaroslav Berdnikov, Vladislav Borisov, Dmitry Kotov, Iurii Mitrankov	HHJ, PLHF, Cu+Cu_200GeV, Cu+Au_200GeV, Run-5, Run-12, single high pT particles, phi	an1402.01 draft an065 an1010 an1374 an770 an911 an964 PPG148
				HHJ, PLHF, Au+Au_200GeV,	an1401.01 draft

DAP Challenges in PHENIX

- Software and infrastructure documentation gradually becoming obsolete (or broken)
  - use of private directories to contain parts of documentation and software leads to loss of data
- Personnel (leaving, graduating or migrating to other projects e.g. *sPHENIX*)
  - continuity of expertise is a problem
  - insufficient effort available to document the software
- Need active management of the analysis software
  - revision control, packaging and archiving of the analysis software developed mainly by individual researchers
- Absence of a comprehensive Conditions Database in PHENIX
  - much of the conditions-type of data in files outside of DB (accessed through scripts)
  - Also, custom “data artifacts” created in individual analyses e.g. geo, augmented dead channel maps/efficiency maps which are hard to trace and reproduce

Current work in PHENIX aligned with DAP

- Looking at embedding procedures in a few types of analyses as examples of complex workflows involving both MC and reco, thus being prime candidates for preservation
- Capturing user/analysis code in custom repos (used as scratch space for now)
  - making a few changes to reduce or eliminate extraneous dependencies, cleanup etc
  - (re) creating documentation
- Often requires reverse engineering
- The interim goal is to create “self-contained” examples of analysis that are essentially ready for preservation
  - i.e. collections of software, data products, documentation, containers and other components as required

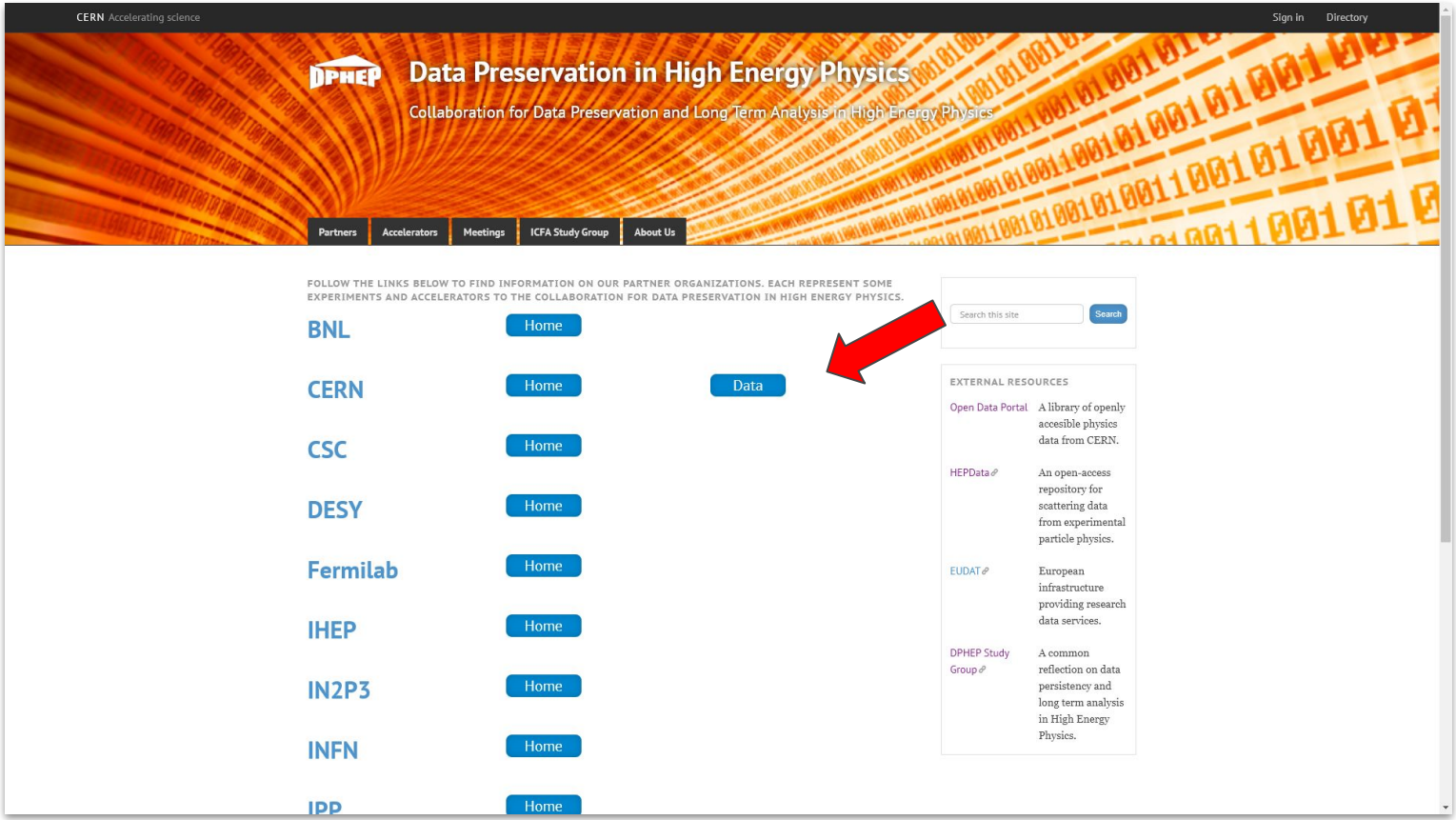
Organizing DAP in PHENIX

- A DAP Task Force has been created, meetings are held periodically
  - includes the documentation, infrastructure, MC and analysis coordinators for PHENIX, with support from the Nuclear and Particle Physics Software Group (NPPS)
- An initial DAP plan has been drafted (September 2019)
- Aiming to join and leverage the community experience - see next slide

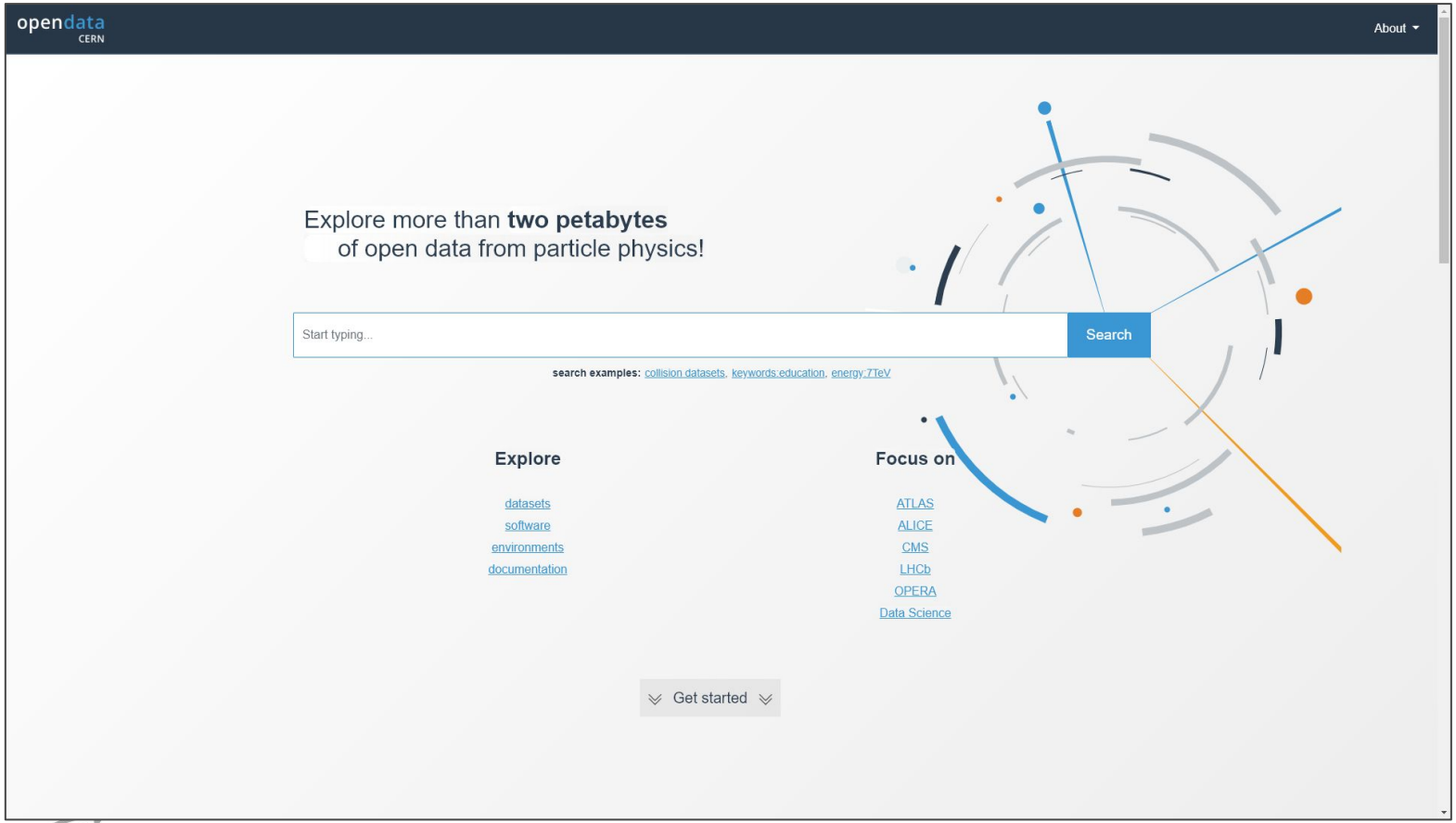
DAP: the community

- The Task Force members and a SDCC representative participated in the DAP workshop at CERN in November 2019 which included the LHC and RHIC experiments
  - The community is being revitalized
  - NB. Almost all content managed on CERN portals is “Tier-1” e.g. Open Data
- Leveraging the expertise of NPPS members, learning from other experiments
- Plans are being developed for a PHENIX Simulations Workshop aiming to survey and preserve existing practices and identify problem areas
- BNL SDCC is an official member of the cooperative DAP tools development and testing effort (a collaboration of ~10 institutions) - next slide
- **The PHENIX DAP effort can serve as a valuable test bed for tools and practices for other experiments (at present STAR, plus sPHENIX going forward)**

DPHEP (the Collaboration)

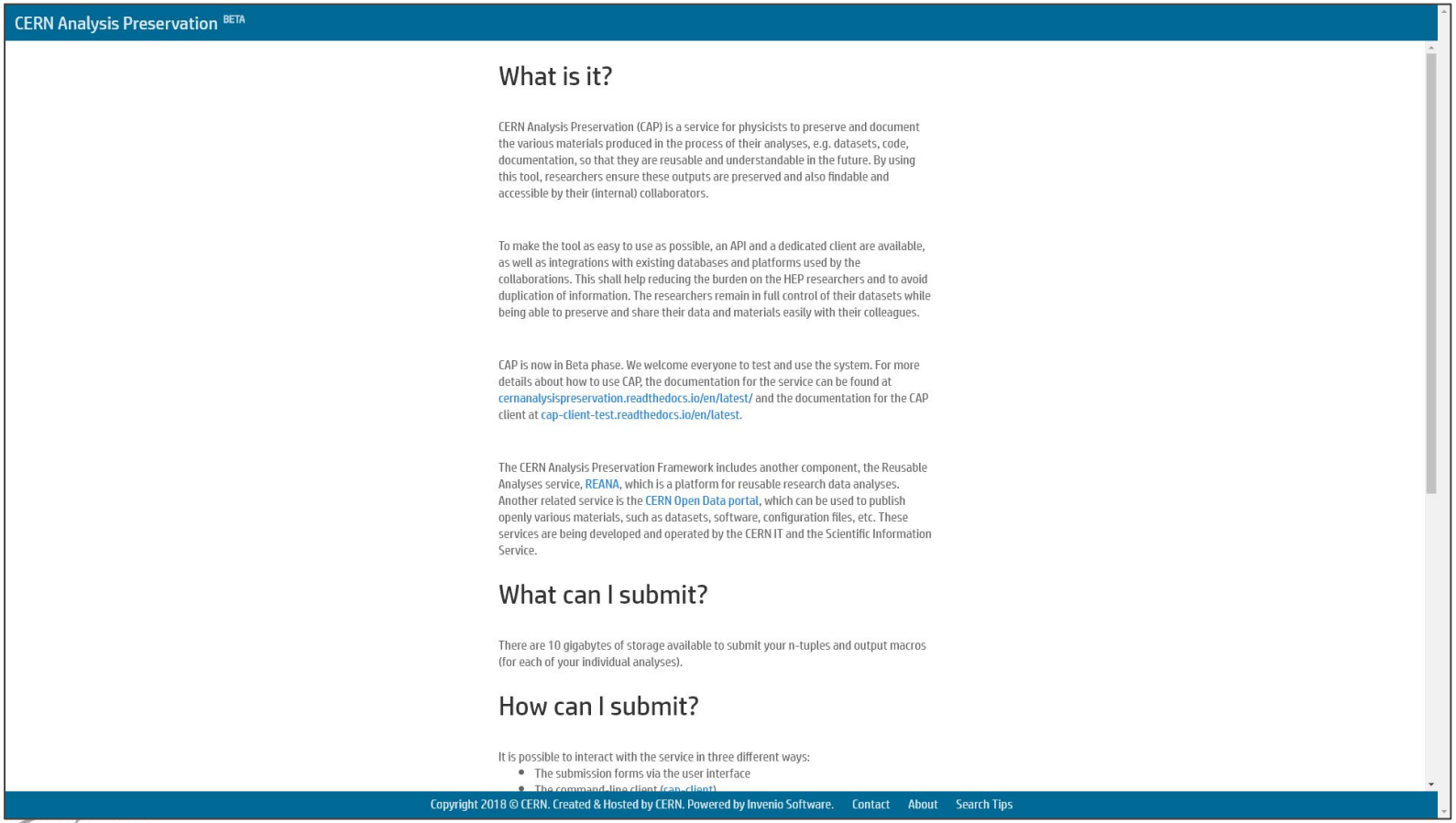


CERN Open Data

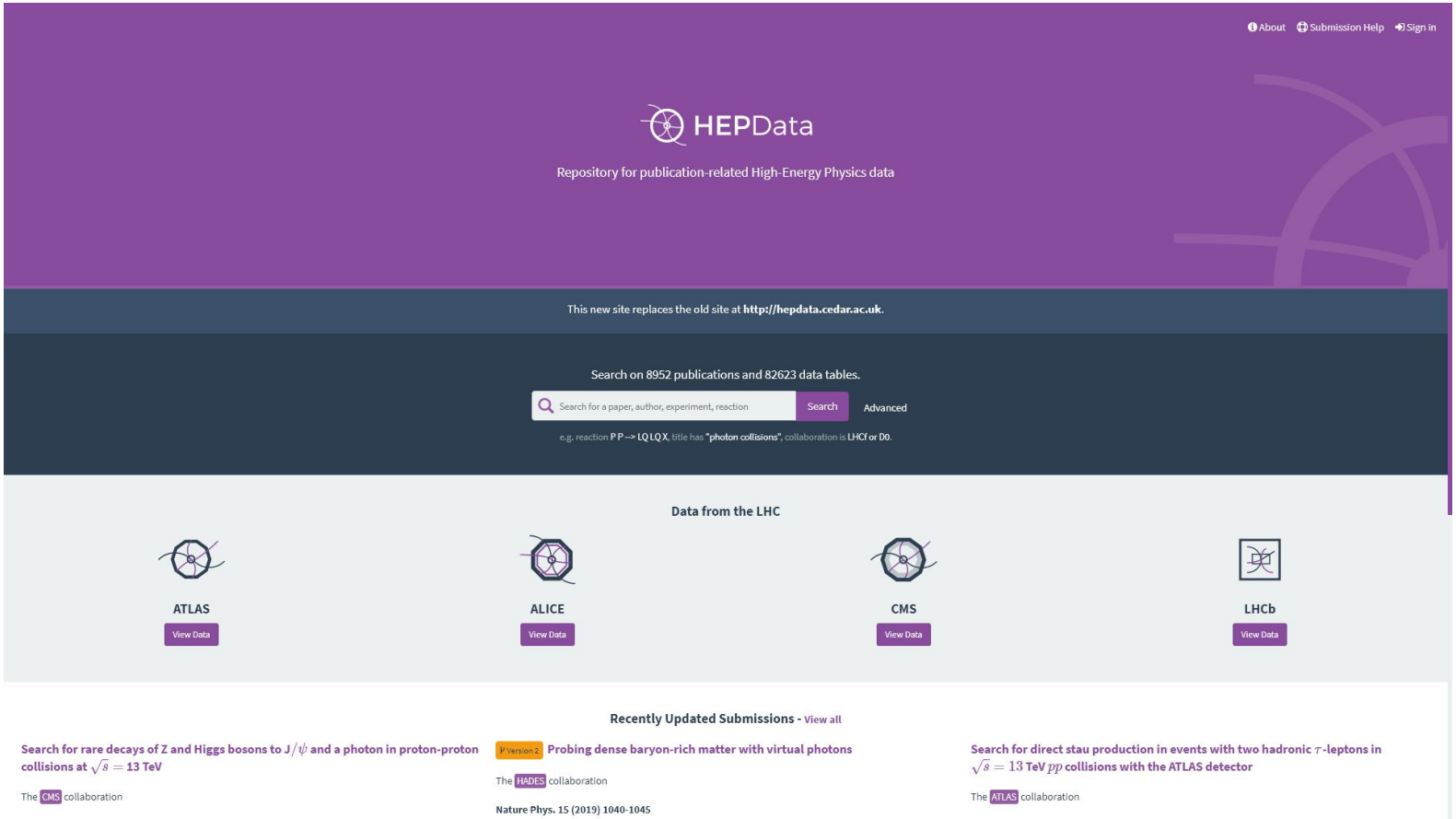


CERN Open Data - sample content

CAP: CERN Analysis Preservation (beta)



The HEPData Portal



Recent trends in DAP

- Use of containers
- Declarative analysis and related toolkits/services
  - Switch from procedural to declarative description of the analysis logic
    - Highly condensed formalism - e.g use YAML or similar/derived format to express the workflow
    - Enforces a good level of software organization and functionality
  - Has the potential to increase the efficiency of analyses; natural fit for CI
  - **REANA service**: “Reproducible research data analysis platform” (see backup slides)
    - Crucially, container-based
- Jupyter notebooks
- New generation of Digital Repository services for Knowledge Management
  - **Invenio** has been the backbone of many CERN-based services e.g. document handling systems (also cf. portal screenshots above)
  - Evolves to include more application areas

Considerations for Invenio

- A proven Digital Repository (CERN), development ongoing (see CERN sites mentioned above)
- Know-how can be preserved in the form of collections of documents, data and software
  - Analysis notes (see above) - hopefully with step-by-step instructions
  - Associated analysis code (in repos or archives e.g. tarballs)
  - Reference data samples, conditions and calibrations type data
  - Containers
- As mentioned above, PHENIX has a custom archive system for analysis notes. There is also a common tool for document management - the “DocDB” developed at FNAL in early 2000s, deployed at FNAL, BNL etc
  - Some useful functionality...but getting old and hard to develop and maintain (Perl)
  - No full text search
- Need a modern and durable tool for DAP, store collections of related objects
  - Full text search can make a big difference
- Migration to a modern platform like **Invenio RDM** would be welcome (RDM is still work in progress)
  - Also an option to use Invenio-based applications currently being developed at BNL (see backup)

Summary of the current DAP plan in PHENIX

- DAP Elements included in the current plan:
  - bit preservation (covered by the facility)
  - Knowledge Management (KM): information gathering + documentation
  - a DAP Metadata system for KM (Invenio the current prime candidate)
  - software (containers, CVMFS for provisioning) + reference data for validation
- Prioritization of the data for DAP based on statistics, stability and quality
- Time profile and distribution of the effort
  - with people leaving the experiment the expertise is dissipating fast
  - there is a window of opportunity to capture knowledge and it must be done **now**
  - resources would be spent most efficiently within 1-2 years from now, whereas in 3+ years the return on investment will be much less

Key Work Areas

- ...excluding Bit Preservation; see Takahito's presentation for wider scope
- Knowledge Management Liaison
  - Curation/review of analysis procedures and data, works with researchers to capture the analysis know-how, goal is to increase engagement on the part of PWGs
- Core software maintenance
- Documentation
- Migration to an enhanced Digital Repository/Metadata/KM system
  - ...e.g. Invenio-based tools
  - Goal is to have well documented and self-contained analysis use cases that are reproducible, which may include samples of reference intermediate data
  - Goes well beyond the current scope of the analysis note template

Request and Deliverables

- DAP effort needed (see Takahito's slides)
  - 0.5 FTE for DAP-specific activity, knowledge curation
  - Contingent on overlapping work areas presented in Takahito's slides (e.g. documentation, software) w/o which DAP won't be possible
- Deliverables
  - A Knowledge Management Digital Repository for DAP materials
  - Rebuilt documentation covering KM in areas crucial for DAP
  - A portfolio of reproducible analyses

Summary

- PHENIX: close correlation and overlap between maintaining analysis capabilities in short and medium term, and longer term DAP
- Initial plans for DAP have been developed, more detailed technical plans in the works
- PHENIX needs added effort in the DAP area
- Participation in the DAP community at large is important (cf. tools and practices)

Backup slides

The Scope

- DAP is not a monolithic category
  - Overlaps with and indeed includes the Open Data access and preservation policies and tool
  - cf. refined data products open to the public (referred to as Tier-1)
  - The de-facto standard for Data Preservation includes 4 Tiers, from raw data (T4) to final data products (T1)
- Focus on Tier-4 as an urgent priority
  - Raw data, infrastructure and tools to process and analyse it

Invenio@BNL

- The SDCC is working with CERN and 10 other multidisciplinary and commercial institutions to build a research data management platform named InvenioRDM
  - Intended to be completed by mid 2020
- Two BNL scientific communities are developing and using Invenio V3 base custom applications hosted by SDCC (full cycle from dev to deployment)
  - open science repository designed to manage/harvest material science related records
  - deployment of Federated Identity to access resources
- Nonproliferation and National Security Department
- ...early adoption by sPHENIX

InvenioRDM: a turn-key open source research data management platform

Lars Holm Nielsen Apr 29, 2019 invenio

CERN has partnered with 10 multidisciplinary institutions and companies to build a turn-key open source research data management platform called InvenioRDM, and grow a diverse community to sustain the platform.

The InvenioRDM project is funded by the CERN Knowledge Transfer Fund, as well as all the participating partners, including:

- Brookhaven National Laboratory (US)
- Caltech Library (US)

REANA

There is only thing that remains in order to make it runnable on the REANA cloud; we need to capture the above structure by means of a `reana.yaml` file:

```
version: 0.4.0
inputs:
  files:
    - code/mycode.py
    - data/mydata.csv
  parameters:
    myparameter: myvalue
workflow:
  type: cwl
  file: workflow/myworkflow.cwl
outputs:
  files:
    - results/myplot.png
```

This file is used by REANA to instantiate and run the analysis on the cloud.

3.7. Declare necessary resources

You can declare other additional runtime dependencies that your workflow needs for successful operation.

CVMFS

If your workflow needs to access CVMFS filesystem, you should provide a `cvmfs` sub-clause of the `resources` clause that would list all the CVMFS volumes that would be mounted for the workflow execution. For example:

```
workflow:
  type: serial
  resources:
    cvmfs:
      - fcc.cern.ch
  specification:
    steps:
      - environment: 'cern/slc6-base'
        commands:
        - ls -l /cvmfs/fcc.cern.ch/sw/views/releases/
```

Kerberos

If your workflow requires Kerberos authentication, you should add `kerberos: true` for the steps in need. Please note that step's docker image (e.g `environment: 'cern/slc6-base'`) should have Kerberos client installed and you should upload keytab file for the Kerberos authentication to work.

Serial example:

v: latest

Declarative Analysis Tools

- Robust investment in experiments like CMS (REANA service running at CERN)
  - NB. between 3 and 4 FTE are involved
- General consensus that this is a good practice… However:
  - Concern 1: prioritization of effort. If the software is ready to be wrapped into a declarative platform, it must be already in a very good shape. Are we there yet?
  - Concern 2: what is the learning curve? Does it fit into the limited resources allocated to DAP in PHENIX?