

# **Data and Analysis Preservation tools for PHENIX**

M.Potekhin

*Nuclear and Particle Physics Software Group*

PHENIX PWG Conveners' Meeting

BNL

March 11th, 2020

# Data and Analysis Preservation

- We'll use the term *DAP* to refer to data **and** analysis preservation
  - cf. data which can't be analyzed is useless
- DAP is commonly described as a union of
  - Bit preservation
  - Software Infrastructure and application code
  - Analysis know-how (knowledge management)
- The goals
  - to have a reproducible analysis capability over a **long period of time**
  - implied: a **modified** or new analysis within the same framework
- There are obvious synergies with maintaining short-term reproducibility - which is required to ensure adequate quality of the analyses
- Accordingly, we are looking at solutions to cover all these items

# DAP: recent history in PHENIX

- Funding agencies in the US increasingly require both new and **existing** experiments to plan/to have DAP capabilities
  - the problem: nobody in the research community has “spare cycles” while extra funding is scarce or non-existent
  - but we should still make an effort
- PHENIX has participated in the DAP workshop at CERN in late 2019
- A Lab review took place where DAP was one of the focus items
  - PHENIX DAP effort is considered a pilot project for other experiments
- A DAP Task Force has been created last year (although available effort is limited)
- In the past few months we have identified a few tools and approaches to help PHENIX address some of the DAP-related work items

# DAP assets in PHENIX

- Bit preservation is taken care of (SDCC) with one tape copy of the data
- Calibration/Conditions type data reside on disk and in some cases under version control
- Production is ongoing
- Many active analyses under way, with O(10) papers published annually
- Well maintained general software infrastructure
  - build, validation and release procedures in place (thanks Chris!)
  - production jobs run in containers etc

# DAP challenges in PHENIX

- But there are also challenges -
  - Software infrastructure documentation becoming (or already being) obsolete
  - Personnel (leaving or migrating to other projects)
  - Continuity of knowledge therefore is a problem (and the software effort underfunded)
  - Lack of comprehensive conditions DB in PHENIX (some data are in DB, some not) requires extra expertise from the researchers (cf. the use of custom “data artefacts” created in individual analyses)
  - Less than strict adherence to the analysis note guidance and template
    - historically, not every note contains 100% of information on reproducible steps/recipes/data products

# DAP = Knowledge Management

- (Long-term) knowledge management is the key to DAP
- Keeping record of minute (but still crucial) details of the analysis procedures requires a non-trivial extra burden on the researcher
  - while getting a quality paper published is the top priority, this usually takes the back seat
- “Foundation software” e.g. PISA and fun4all need an updated and reliable documentation with working tutorials so researchers are in control of their software both in the short and in the long term
  - a review of the tutorials has started
- A lot of the “know how” is folded into the analysis software which needs to be reviewed and documented
  - workflows can be relatively complex
- There is a fairly narrow window of opportunity to address at least some of the challenges

# DAP topics for today

- Focus on options and existing solutions suitable for DAP in PHENIX:
  - Curation and consolidation of mostly static documentation on a newly designed website based on a proven technology
  - A possibility to leverage an existing and powerful platform for managing papers and similar documents
- We'll leave specific current software issues for a discussion at a later date

# PHENIX Web documentation

- Currently the information describing the detector, run conditions, foundation software and other topics is often fragmented and sometimes outdated (and sometime woefully obsolete)
  - e.g. kept on separate servers
  - Broken links can be really annoying
  - Mix of HTML pages, Wikis and other platforms
  - A lot of information is no longer relevant for DAP so it dilutes the content
- A large fraction of the documents exist as HTML which is difficult to edit and makes keeping consistency of references a challenge
- A large part is outside of version control
- Fixing existing websites is not practical for a number of reasons


# The New PHENIX Website

- To better serve the short and long-term goals of PHENIX the Collaboration needs a consolidated and curated Web resource
- Running a content management system (such as Wiki, Drupal etc) over a long period of time is possible
- ...but it is not optimal:
  - the material is static anyway so a database server is an overkill as a resource
  - it would require DB and other support, security updates, admin expertise etc
  - handling structured data is possible but not straightforward (see below)
  - version control is less than ideal
  - backup is possible but not straightforward
- Ideally, the long-term Web platform would combine the following:
  - static content (to simplify long-term maintenance and for better security)
  - structured data kept separately from layout (for referential integrity)
- Luckily, ready solutions do exist. Examples are in the following slides.


# Example 1: HEP Software Foundation


- <https://hepsoftwarefoundation.org/>

[HEP Software Foundation](#) [Activities](#) [Working Groups](#) [Communication](#) [Projects & Support](#) [About](#)



The HEP Software Foundation facilitates cooperation and **common efforts** in High Energy Physics software and computing internationally.

 **HSF and WLCG Workshop, Lund 11-15 May 2020** ([more info](#))

 **PyHEP 2020 "Python in HEP" Workshop, Austin, Texas, USA, 11-13 July 2020** ([more info](#))

## Meetings

The HSF holds **regular meetings** in its activity areas and has weekly coordination meetings as well. All of our meetings are open for everyone to join.


- HSF Weekly Meeting #181, 27 February, 2020
- HSF Software Tools and Packaging Working Group Meeting #1, 26th February 2020
- HSF Weekly Meeting #180, 13 February, 2020

[Upcoming HSF events »](#)

[Full list of past meetings »](#)

## From JLab to Lund

The last HSF Workshop, which we held in JLab, was a **great success**. Now we continue the series, joining with WLCG again for the **next workshop** hosted by our colleagues in Lund.



We have the pleasure of being in the beautiful main building of the university and we look forward to another very productive workshop.

*Registration now open!*

## Activities

We organise many activities, from our **working groups**, to organising **events**, to supporting projects as **HSF projects**, and helping communication within the community through our **discussion forums**, **technical notes** and a **knowledge base**.

[How to get involved »](#)


Thanks to [GitHub Pages](#), [Jekyll](#) and [Bootstrap](#)

## Example 2: NPPS Group at BNL

- <https://npps.bnl.gov/>

BNL NPPS
Experiments and Projects ▾
Tech Teams ▾
Software ▾
Events ▾
Resources ▾
About ▾

---




**BROOKHAVEN**  
NATIONAL LABORATORY  
NUCLEAR AND PARTICLE PHYSICS  
SOFTWARE GROUP

The Nuclear and Particle Physics Software (NPPS) Group in Brookhaven National Laboratory's Physics Department participates in a wide range of experiments across BNL's nuclear and particle physics programs. NPPS provides software and expertise across many technical areas, with a particular emphasis on common software solutions.

## News

[More info](#)

 **BNL/JLab S&C Round Table on Jupyter**

## Experiments & Projects

NPPS members participate in many of BNL's nuclear and particle physics experiments and programs, working as members of experiment software teams. We also develop and participate in R&D projects, and community projects such as the HSF.

- ATLAS at the LHC (CERN)
- Belle II at SuperKEKB (KEK)
- DUNE Long Baseline Neutrino Facility (FNAL & Sanford)
- Electron Ion Collider (EIC)
- Google-ATLAS HL-LHC Project
- HEP Software Foundation (HSF)
- Lattice QCD
- Large Synoptic Survey Telescope (LSST)
- PHENIX at RHIC (BNL)
- sPHENIX at RHIC (BNL)
- STAR at RHIC (BNL)

## Technical Teams

Beyond experiment-specific work, NPPS members and collaborators make up technical teams spanning the group's many areas of expertise. Teams may be actively engaged in cross-experiment projects or may act as technical forums sharing expertise and exploring common project opportunities.

- Analysis tools
- Collaborative tools, documentation and training
- Core software & advanced algorithms
- Databases
- Data and analysis preservation
- Distributed data management
- Event data storage and I/O
- High performance computing
- Real-time analysis
- Reconstruction
- Simulation

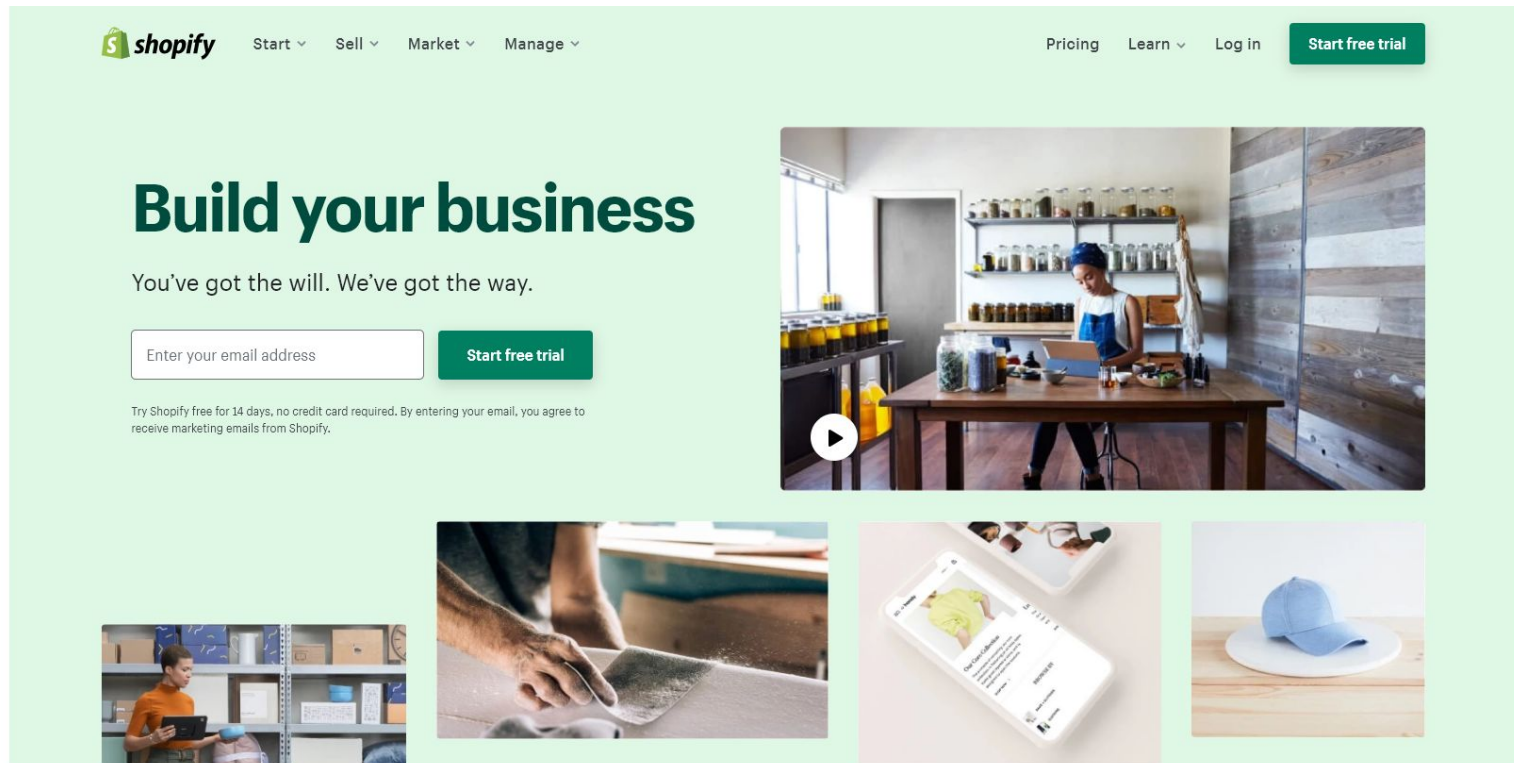
## Software

NPPS members and collaborators have developed and/or have expertise on many software products and projects, many of which are (or could be) multi-experiment **common software**.

- ACTS**
- Athena**
- ATLAS CI and nightly system
- Conditions database**
- Data Carousel
- DIRAC**
- DUNE prompt processing system and Data Quality Monitor
- dunetpc
- EicRoot**
- eic-smear**
- ATLAS event service
- Fun4All**

# Example 3: Shopify (e-commerce)

- <https://www.shopify.com/>

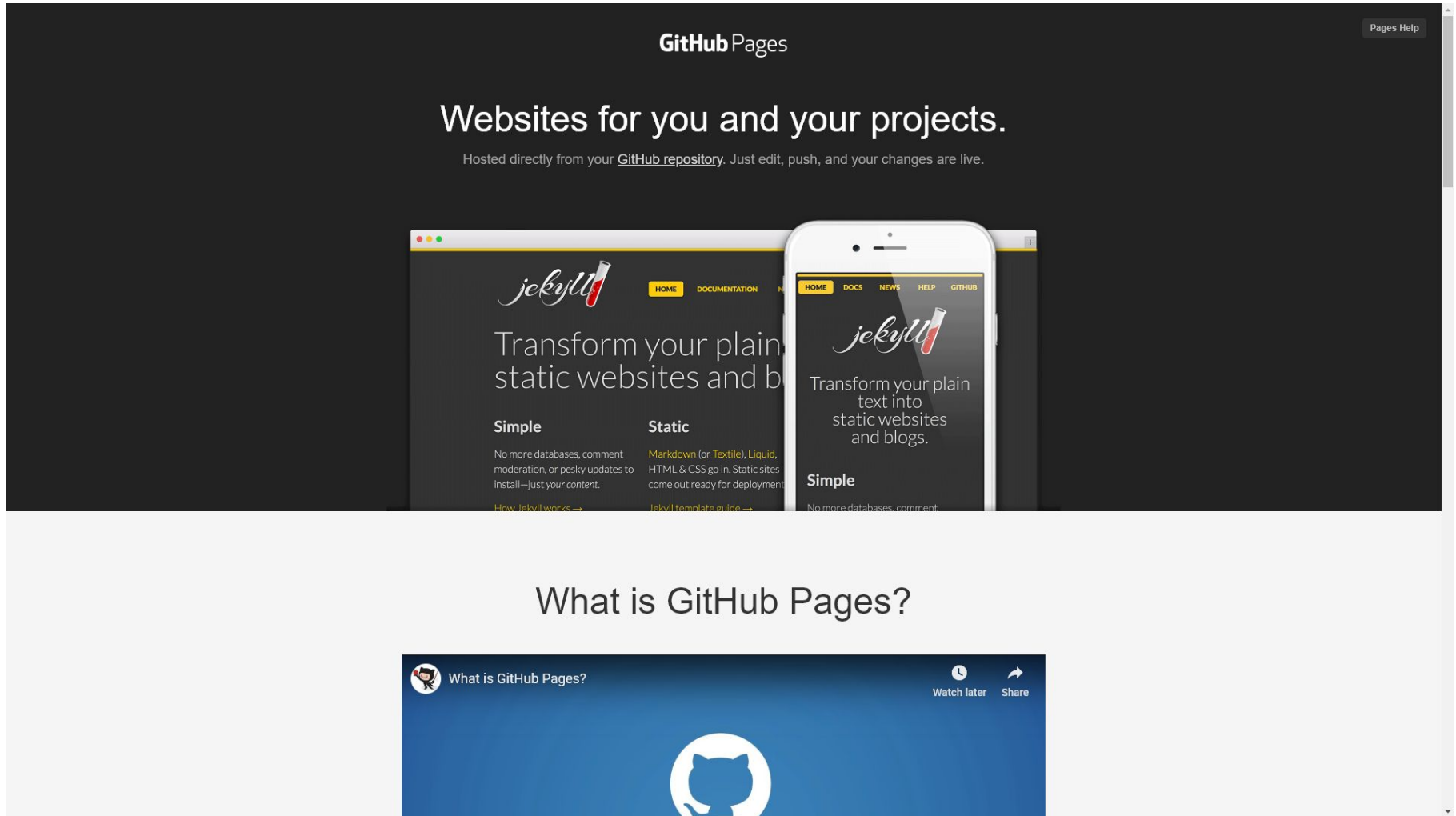


With you from first sale to full scale

One platform with all the ecommerce and point of sale features you need to start, run, and grow your business.

# Example 4: GitHub pages

- <https://pages.github.com/>



# The platform

- Common in all the examples given above is the static website generator called “Jekyll” which utilizes a standard template language “Liquid” for layouts, parses Markdown syntax for content and YAML for the data component
  - YAML is essentially a human-readable variant of JSON
  - Likewise, Markdown is a lot more readable and easier to format than HTML
  - “Liquid” is quite similar to other popular template languages and easy to learn
- YAML can be embedded in the individual pages as needed
- A lot of flexibility in how the data is referenced
- This site generator is fully integrated into GitHub
  - one just has to configure a repository in a compatible way
  - site is rebuilt with every push to GitHub
  - ideal platform for collaborative development with state-of-the-art version control
- Once generated, the site is a hierarchy of a few folders and is portable to any web server and even accessible through a “file:///” handle in your local browser
  - i.e. you don’t even need a web server at all, the whole website can be shipped through e-mail, Dropbox or even USB stick
- At the same time, GitHub provides hosting of such sites for free and we expect it to remain in business on a time scale relevant for the PHENIX data preservation mandate
  - also, fairly low effort required to deploy such a site at BNL or elsewhere

# The plan

- The goal is to create an optimal experience for the user looking for information about the experiment, its detector subsystems, foundation software and hopefully analysis know-how
  - NB. theses and analysis notes will be kept in a dedicated document management system as it is a different domain - see the second part of this presentation
- Look at best practices and borrow design elements from other projects
- Keep Web development under the official PHENIX Collaboration umbrella on GitHub
  - <https://github.com/PhenixCollaboration>
- Create a basic layout for the pages and also for the data on disk (almost done)
- The most difficult part: survey, curate, improve and organize various materials and make them ready for inclusion on the new site
- Status:
  - a prototype website has been stubbed out recently
  - very little content right now but keep growing continuously
  - thanks to Gabor for collecting materials
  - your input is needed and appreciated
  - please help in gathering and reviewing materials
- This is the current location: <https://phenixcollaboration.github.io/web/>

# The landing page

- <https://phenixcollaboration.github.io/web/>



PHENIX, the Pioneering High Energy Nuclear Interaction eXperiment, is an exploratory experiment for the investigation of high energy collisions of heavy ions and protons. It is the largest of the four experiments that have taken data at the Relativistic Heavy Ion Collider. Data-taking was finished in 2016 and the PHENIX Collaboration is now analyzing large data samples collected, prioritizing those with a unique physics reach.



# Summary of runs

- <https://phenixcollaboration.github.io/web/experiment/runs.html>

PHENIX Home			
	The Experiment ▾	Detectors ▾	Foundation Software ▾
			Analysis ▾
			About ▾
Summary of Runs			
Run	Period	Coordinator(s)	
<a href="#">Run 01</a>	1999-2001	Achim Franz, BNL.	
<a href="#">Run 02</a>	2001-2002	Anthony D. Frawley, FSU.	
<a href="#">Run 03</a>	2002-2003	Matthias Grosse Perdekamp, UIUC and RBRC.	
<a href="#">Run 04</a>	2003-2004	Terry C. Awes, ORNL.	
<a href="#">Run 05</a>	2004-2005	John Lajoie, ISU.	
<a href="#">Run 06</a>	2005-2006	Abhay Deshpande, SUNYSB and RBRC.	
<a href="#">Run 07</a>	2006-2007	Michael J. Leitch, LANL.	
<a href="#">Run 08</a>	2007-2008	Michael J. Leitch, LANL.	
<a href="#">Run 09</a>	2008-2009	John Haggerty, BNL.	
<a href="#">Run 10</a>	2009-2010	Stefan Bathe, RIKEN-BNL.	
<a href="#">Run 11</a>	2010-2011	Takao Sakaguchi, BNL.	
<a href="#">Run 12</a>	2011-2012	Xiaochun He, GSU.	
<a href="#">Run 13</a>	2013	Hubert van Hecke, LANL.	
<a href="#">Run 14</a>	2014	Klaus Dehmelt, SUNYSB.	
<a href="#">Run 15</a>	2015	Douglas Fields, UNM; Itaru Nakagawa, Riken/RBRC (spin).	
<a href="#">Run 16</a>	2016	Denis Jouan, IPNO.	

## Summary tables (click for larger image)

Runs 01-08 summary table

TABLE 1.1. Summary of RHIC Luminosity 2001-2008

Run	Year	Species	$\sqrt{s_{NN}}$ (GeV)	Integrated Luminosity	$N_{events}$	Size of Data Set
01	2000	Au-Au	130	$1b^{-1}$	10M	3TB
		Au-Au	200	$24b^{-1}$	10M	10 TB
02	2001/02	Au-Au	19		<M	
		p-p	200	$0.15pb^{-1}$	3.7B	20TB
03	2002/03	d-Au	200	$2.74nb^{-1}$	5.5B	46 TB
		p-p	200	$0.35pb^{-1}$	6.6B	35 TB
04	2003/04	Au-Au	200	$241\mu b^{-1}$	1.5B	270 TB
		Au-Au	62.4	$9\mu b^{-1}$	58M	10 TB
		Cu-Cu	200	$3nb^{-1}$	8.6 B	173 TB
05	2005	Cu-Cu	62.4	$0.19nb^{-1}$	0.5 B	48 TB
		Cu-Cu	22.4	$2.7\mu b^{-1}$	3.7B	20TB
		p-p	200	$3.8pb^{-1}$	85 B	262TB
06	2006	p-p	200	$10.7pb^{-1}$	233B	310 TB
		p-p	62.4	$0.1pb^{-1}$	28B	25 TB
07	2007	Au-Au	200	$813\mu b^{-1}$	5.1B	650 TB
		d-Au	200	$80nb^{-1}$	160 B	437 TB
08	2007/08	p-p	200	$5.2pb^{-1}$	115 B	118 TB
		Au-Au	9.2		< 5k	

Runs 09-16 summary table

PHENIX Data sets 2009-2016 (Runs 9-16)

The Table below summarizes general indications about data sets collected by PHENIX in Runs 9-16.

Run	Species	$\sqrt{s_{NN}}$ (GeV)	Luminosity (nb <sup>-1</sup> /s)	PHENIX Data Set	File	Comments
9	p-p	200	11.6B	11.6B	11.6B	
	p-p	200	20.6B	20.6B	20.6B	
10	d-Au	200	1.7B	1.7B	1.7B	Shimizu risk
	d-Au	62.4	0.11B	0.11B	0.11B	
	d-Au	34	0.01B	0.01B	0.01B	
	d-Au	1.7	0.1B	0.1B	0.1B	
11	p-p	200	18.6B	18.6B	18.6B	PHENIX Data Set
	d-Au	19.4	1.1B	1.1B	1.1B	PHENIX Data Set
	d-Au	200	1.7B	1.7B	1.7B	PHENIX Data Set
	d-Au	37	1.1B	1.1B	1.1B	PHENIX Data Set
12	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
13	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
14	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
15	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
16	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set
	p-p	200	10.6B	10.6B	10.6B	PHENIX Data Set

# Run configuration gallery

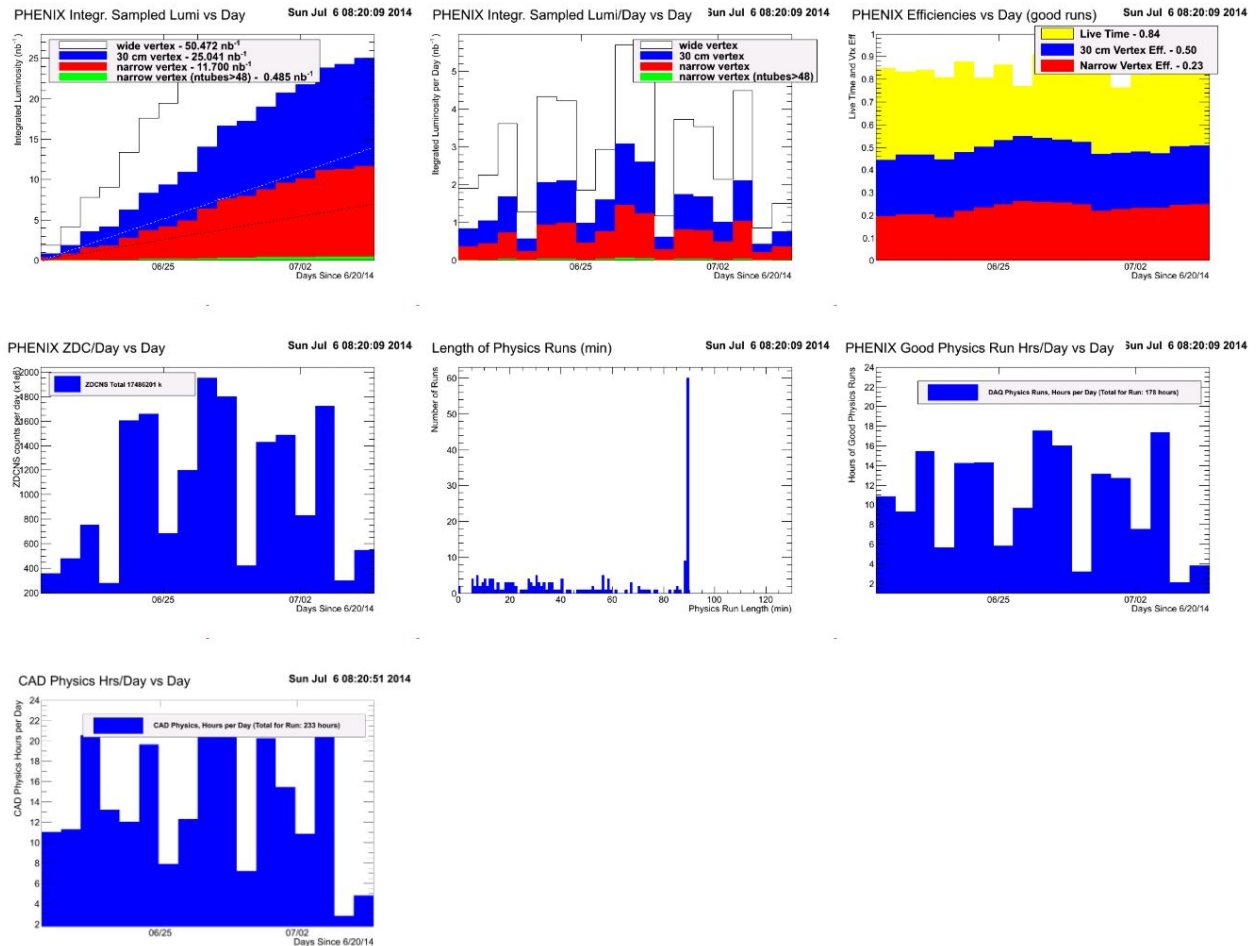
- [https://phenixcollaboration.github.io/web/detectors/run\\_configuration\\_gallery.html](https://phenixcollaboration.github.io/web/detectors/run_configuration_gallery.html)



# Example of luminosity plots

- <https://phenixcollaboration.github.io/web/runs/run14.html>

Run 14 Luminosity accumulation and related data



# Example of a page layout

- [https://phenixcollaboration.github.io/web/detectors/central\\_arm.html](https://phenixcollaboration.github.io/web/detectors/central_arm.html)

PHENIX Home

The Experiment ▾

Detectors ▾

Foundation Software ▾

Analysis ▾

About ▾

## Central Arm Detectors

Name	Role
<a href="#">Drift Chamber</a>	Measures the position and momentum of charged particles
<a href="#">Pad Chambers</a>	Measures the position of charged particles with precision
<a href="#">Ring Imaging Cherenkov</a>	Identifies Electrons
<a href="#">Hadron Blind Detector</a>	Identifies electrons without detecting hadrons
<a href="#">Time Expansion Chamber</a>	Measures the position and momentum of charged particles. Identifies particles.
<a href="#">Time-of-Flight</a>	Measures the position of charged particles. Identifies particles.
<a href="#">Aerogel Cerenkov Counter</a>	Identifies particles at high transverse momentum.
<a href="#">Time-of-Flight West</a>	High resolution particle identification.
<a href="#">Electromagnetic Calorimeter</a>	Measures the position and energy of charged and neutral particles. Identifies photons and charged particles.
<a href="#">Silicon Vertex Tracker</a>	Charged particle tracking near the collision vertex.

# ...and the code

Files for individual detectors are picked up automatically (according to the category) from a folder and sorted according to the weight assigned, then the page is automatically formed.

Same code is used for all detector categories.

```
# {{ page.title }}
<table WIDTH="100%">
<tr><th>Name</th><th>Role</th></tr>
{% assign items = site.detectors | sort: 'weight' %}
{% for detector in items %}
{% if detector.category == page.detector_category %}
{% include detector_category.md content=detector %}
{% endif %}
{% endfor %}
</table>
```

# YAML

## Example 1:

### RUN SUMMARY (will include reference to luminosity data etc)

- run: 01  
period: 1999-2001  
coordinator: Achim Franz, BNL.
- run: 02  
period: 2001-2002  
coordinator: Anthony D. Frawley, FSU.

## Example 2:

### MAIN PHENIX PHOTO GALLERY

- path: /images/photos/3177203852\_d3d30ae925\_k.jpg  
title:  
gallery: main  
type: photo
- path: /images/photos/3253561842\_6960470407\_k.jpg  
title:  
gallery: main  
type: photo

# The core issue

- It is up to the conveners and other members of the PHENIX leadership to define the following:
  - WHAT information should be stored and to what depth
  - HOW it should be organized and cross-referenced
  - WHO is willing to contribute the content and perform quality assurance of information published on the site
- This is the core issue and all and any technology decisions are less important.
- Please visit the prototype site. Let us take a look at a few more screenshots in the following slides.

# ...back to the site: the “experiment” menu

PHENIX Home

The Experiment ▾

Detectors ▾

Foundation Software ▾

Analysis ▾

About ▾

Concept

History

Summary of Runs

Publications



PHENIX, the Pioneering High Energy Nuclear Interaction eXperiment, is an exploratory experiment for the investigation of high energy collisions of heavy ions and protons. It is the largest of the four experiments that have taken data at the Relativistic Heavy Ion Collider. Data-taking was finished in 2016 and the PHENIX Collaboration is now analyzing large data samples collected, prioritizing those with a unique physics reach.

# The detectors menu

PHENIX Home

The Experiment ▾

Detectors ▾

Foundation Software ▾

Analysis ▾

About ▾

Detectors Overview

PHENIX Photo Gallery

Run Configuration Gallery

Central Arm Detectors

Muon Arm Detectors

Event Characterization Detectors

Magnet



ing High Energy Nuclear Interaction  
laboratory experiment for the investigation  
ons of heavy ions and protons. It is the  
periments that have taken data at the  
n Collider. Data-taking was finished in  
2016 and the PHENIX Collaboration is now analyzing large data  
samples collected, prioritizing those with a unique physics  
reach.

# Another example of a detector category menu

PHENIX Home

The Experiment ▾

Detectors ▾

Foundation Software ▾

Analysis ▾

About ▾

## Event Characterization Detectors

### Name

[Beam-Beam Counters](#)

[Zero Degree Calorimeters and Shower Max Detectors](#)

[Forward Calorimeters](#)

[Multiplicity Vertex Detector](#)

[Reaction Plane Detector](#)

### Role

Measures collision location and centrality. Starts the stopwatch for an event.

Measures collision location and centrality.

For d+Au collisions, it measures surviving neutrons and protons from the original deuteron.

Measures collision location and charged particle distributions.

Accurately measures the reaction plane of each collision.

# The new site: issues

- The site is work in progress
- The data immediately available is incomplete e.g. configuration diagrams for some of the runs are missing; compiling various detector parameters is labor intensive
  - can the conveners help to get this effort going?
- Creating fully functional documentation and tutorial pages for PISA and fun4all will take effort
- Likewise, summarizing common patterns in the analysis techniques used in PHENIX (e.g. embedding)

# Document management

- Historically PHENIX relies on custom PHP applications to manage its analysis notes and theses materials
- These are workable solutions but we need to consider this in the longer run (e.g. maintenance, updates etc)
- Query functions can be better
- Need to revisit what information needs to be “internal” for correct classification going forward
- Alternatives include the “DocDB” system widely used at FNAL and elsewhere and more recently at BNL
  - however that system also shows its age
- All things considered, it is better to use a modern solution with wide community acceptance and which is expected to be supported for a long period of time
- We now look at the (substantial) experience with document management at CERN and in a wider community
  - BNL is an active partner in this activity
  - An integral part of the wider domain of “research data management”

# Document management at CERN

- Many information systems at CERN are based on an array of services rolled out in 2002 and then called CDS, more recently rebranded as “Invenio”
  - a proven platform
  - a recent version of Invenio-based app is actively evaluated in sPHENIX
- The focus of the current development is to transform the system from a service create on site and hosted at CERN to a set of portable products
  - i.e. such that it can be installed at any site
  - due to complexity this takes time
- Next generation: Invenio RDM stands for “Research Data Management” and promises to be a “turnkey” system adaptable to many data and document management needs in many areas including those outside HEP/nuclear physics
  - while LHC is still the main driver and a leading source of funding
- Essentially a digital repository
- BNL is a partner in this development (as a test site)
  - BNL maintains a working relationship with DAP experts at CERN
- For Invenio RDM delivery is expected in late 2020 - and perhaps it will take longer for the system to become stable and ready for prime time
  - may not fit well with the PHENIX timeline
- What can we use now?
  - let’s take a look at the current Invenio-based service at CERN called “Zenodo”

# Zenodo

- <https://zenodo.org/>
- The name derived from Ζηνόδοτος, a Greek scholar and inventor of metadata in 280 BC

The screenshot shows the Zenodo website interface. At the top is a blue header with the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. A user profile for 'potekhin@bnl.gov' is visible. Below the header, a light blue banner contains a notice about a temporary unavailability on Wednesday 11th. The main content area is divided into two columns. The left column, titled 'Recent uploads', lists four items: 1) 'Evaluation of the potential incidence of COVID-19 and effectiveness of contention measures in Spain: a data-driven approach' by Aleta, Alberto and Moreno, Yamir, dated March 9, 2020. 2) 'episanty/RB-SFA: Version 2.2.0' by E Pisanty, dated March 2, 2020. 3) 'HopkinsIDD/ncov\_incubation: Submission code for The incubation period of COVID-19 from publicly reported confirmed cases: estimation and application.' by Stephen A Lauer et al., dated February 28, 2020. 4) 'Gene Ontology Data Archive' by Carbon, Seth and Mungall, Chris, dated July 2, 2018. The right column contains three promotional boxes: 'Zenodo now supports usage statistics!' with a line graph icon, 'Using GitHub?' with the GitHub logo, and 'Zenodo in a nutshell' which lists key features: Research, Shared; Citeable, Discoverable; Communities; Funding; Flexible licensing; and Safe. At the bottom right, there is a 'Tweets by @ZENODO\_ORG' section showing a tweet from the Zenodo account.

# Zenodo in a nutshell

- <https://zenodo.org/>

## Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citeable. Discoverable.** — uploads gets a Digital Object Identifier (DOI) to make them easily and uniquely citeable.
- **Communities** — create and curate your own community for a workshop, project, department, journal, into which you can accept or reject uploads. Your own complete digital repository!
- **Funding** — identify grants, integrated in reporting lines for research funded by the European Commission via OpenAIRE.
- **Flexible licensing** — because not everything is under Creative Commons.
- **Safe** — your research output is stored safely for the future in the same cloud infrastructure as CERN's own LHC research data.

# Zenodo: how do I use it?

- Pretty straightforward - upload any file and dress it up with metadata e.g. tags
  - in that regard similar to the current PHENIX notes system, sPHENIX Invenio app and DocDB (but read below about search capabilities)
- Importantly, access is controlled
  - closed (just the uploader can see)
  - restricted (a user can request access which you then explicitly grant)
  - public (self explanatory)
- Metadata can be edited at any time later but not the source file w/o changing the version
- Version support is important
- Example: in the next slide I show how I uploaded my poster for the CHEP'19 conference and how it is managed

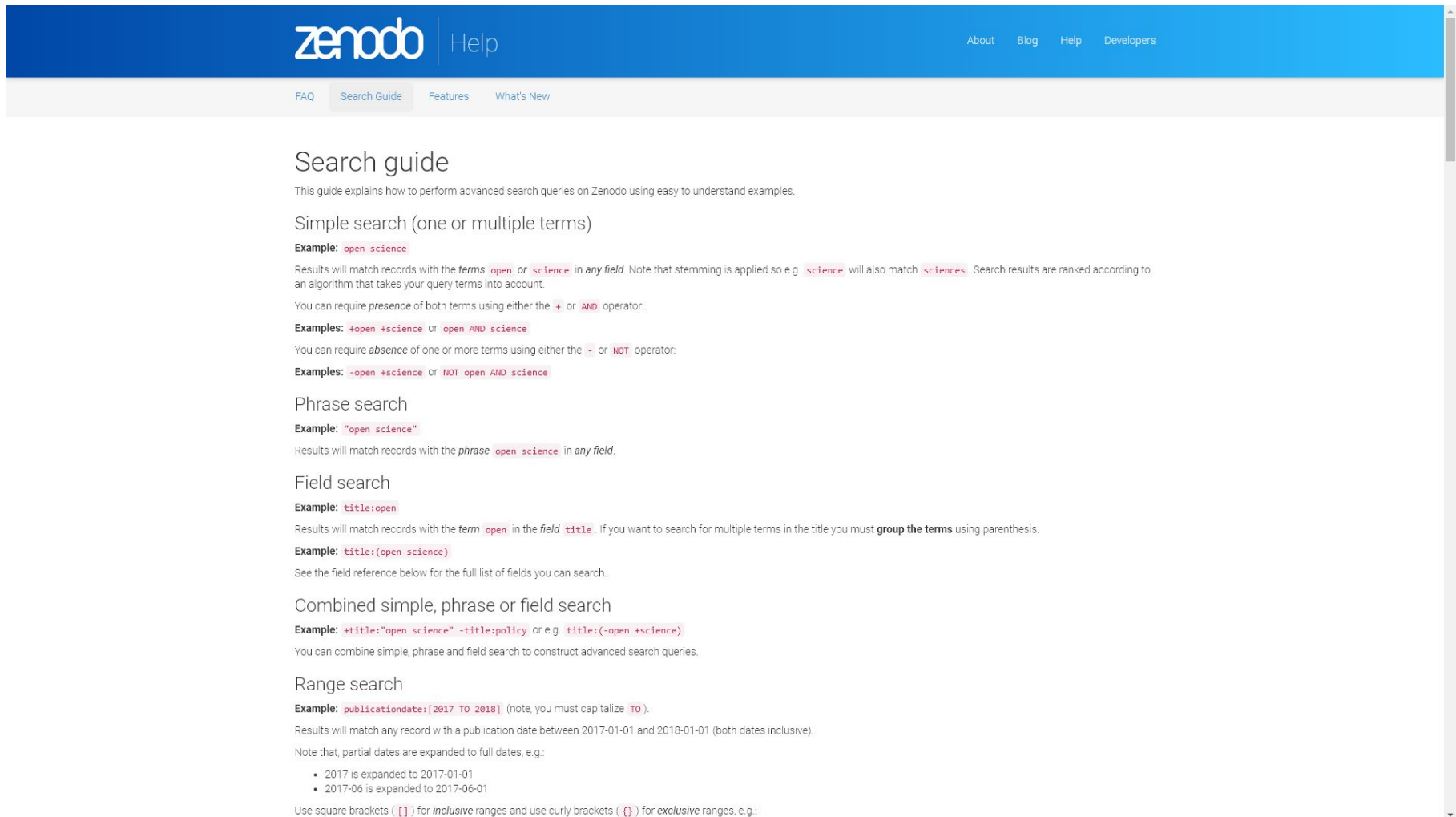
# Zenodo document example

- Note tags, DOI (automatically generated), the conference link, the edit button and the version button

The screenshot shows a Zenodo document page. At the top is a blue header with the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. The user 'potekhin@bnl.gov' is logged in. The document title is 'Evolution of the Data Quality Monitoring and Prompt Processing System in the protoDUNE-SP experiment' by Maxim Potekhin, dated March 2, 2020. It has 18 views and 11 downloads. The document is indexed in OpenAIRE. The publication date is March 2, 2020, and the DOI is 10.5281/zenodo.3693788. The keywords are 'dqm', 'prompt processing', 'data quality monitoring', and 'DUNE'. The meeting is the 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019) in Adelaide, Australia, 4-8 November 2019. The license is Creative Commons Attribution 4.0 International. The document includes a preview of a presentation slide titled 'Evolving monitoring' which shows various plots and text describing the data quality monitoring system. The slide content includes: 'added, modified and removed from the processing scheme according to the needs of the experiment. At various points in time the following payloads were run:', '2D event display based on raw data', 'The TPC monitoring application with an extensive set of histograms at various levels of channel aggregation: FFT, drifts, etc.', 'Health monitor for the Front End Electronics motherboards', 'Prompt reconstruction with metrics such as track candidate count, hit count etc.', 'Track candidate based-argon purity estimators with time series stored in the DB', 'Signal-to-noise ratio monitoring', 'Data preparation for the 3D event display implemented on a separate server', and 'Out-of-band feed from the hardware argon purity monitor'. It also mentions 'Monitoring tools were added to the prompt processing system to address a few infrastructure issues encountered during its operation, e.g. collecting metrics for the transfer latency, batch system performance, storage required by the input data and displaying alarms to operators as necessary.'

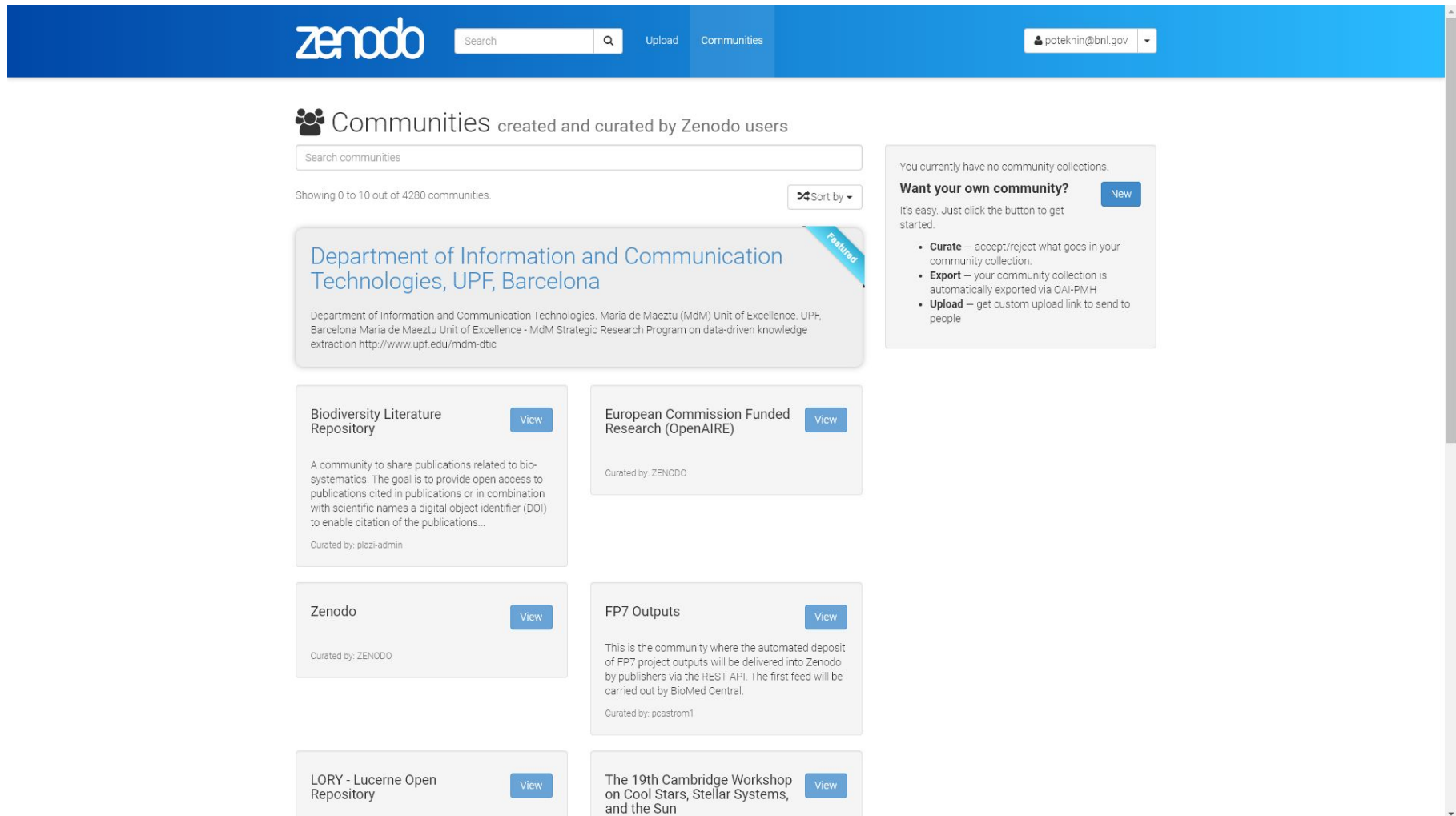
# Zenodo search capabilities

- <https://help.zenodo.org/guides/search/>
- pretty advanced... elastic search (full text) is work in progress for Invenio RDM...

The screenshot shows the Zenodo Help page for the Search Guide. The header is blue with the Zenodo logo and 'Help' text. Navigation links include 'About', 'Blog', 'Help', and 'Developers'. A secondary navigation bar contains 'FAQ', 'Search Guide' (active), 'Features', and 'What's New'. The main content area is titled 'Search guide' and explains how to perform advanced search queries. It covers simple search (one or multiple terms), phrase search, field search, combined simple, phrase or field search, and range search. Examples are provided for each search type, such as 'open science', 'title:open science', and 'publicationdate:[2017 TO 2018]'. The text is formatted with color-coded examples and operators like AND, OR, NOT, and parentheses.

# Zenodo communities

- Users can create communities (e.g. “PHENIX”)
- Not mandatory but might be helpful to maintain a collection of documents specific to PHENIX



The screenshot shows the Zenodo website's 'Communities' section. The header is blue with the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. A user profile 'potekhin@bnl.gov' is logged in. The main content area is titled 'Communities created and curated by Zenodo users' and includes a search bar and a sort dropdown. A featured community card for 'Department of Information and Communication Technologies, UPF, Barcelona' is highlighted with a 'Featured' badge. Below it are several other community cards, each with a 'View' button: 'Biodiversity Literature Repository' (curated by plazi-admin), 'European Commission Funded Research (OpenAIRE)' (curated by ZENODO), 'Zenodo' (curated by ZENODO), 'FP7 Outputs' (curated by pcastrom1), 'LORY - Lucerne Open Repository' (curated by pcastrom1), and 'The 19th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun' (curated by pcastrom1). A sidebar on the right prompts the user to 'Want your own community?' with a 'New' button and lists actions: Curate, Export, and Upload.

zenodo Search Upload Communities potekhin@bnl.gov

Communities created and curated by Zenodo users

Search communities

Showing 0 to 10 out of 4280 communities. Sort by

**Department of Information and Communication Technologies, UPF, Barcelona** *Featured*

Department of Information and Communication Technologies, Maria de Maeztu (MdM) Unit of Excellence, UPF, Barcelona Maria de Maeztu Unit of Excellence - MdM Strategic Research Program on data-driven knowledge extraction <http://www.upf.edu/mdm-dtic>

**Biodiversity Literature Repository** View

A community to share publications related to bio-systematics. The goal is to provide open access to publications cited in publications or in combination with scientific names a digital object identifier (DOI) to enable citation of the publications...

Curated by: plazi-admin

**European Commission Funded Research (OpenAIRE)** View

Curated by: ZENODO

**Zenodo** View

Curated by: ZENODO

**FP7 Outputs** View

This is the community where the automated deposit of FP7 project outputs will be delivered into Zenodo by publishers via the REST API. The first feed will be carried out by BioMed Central.

Curated by: pcastrom1

**LORY - Lucerne Open Repository** View

**The 19th Cambridge Workshop on Cool Stars, Stellar Systems, and the Sun** View

You currently have no community collections.

**Want your own community?** New

It's easy. Just click the button to get started.

- **Curate** — accept/reject what goes in your community collection.
- **Export** — your community collection is automatically exported via OAI-PMH.
- **Upload** — get custom upload link to send to people

# Zenodo: other considerations

- Integration with GitHub which make software code citeable (DOI)
- Service is free
- Information from the developers is such that it should be possible to migrate to a new repository from the same product line later (e.g. when Invenio RDM becomes available at BNL)
- The most important thing - Zenodo presents a superset of functionality implemented in DocDB (the FNAL document management system also adopted at BNL)
  - from experience, this is a good thing
- Community + Access control = Ownership
- Any type of data can be archived (datasets, ntuples) - with some volume limitations
  - this can be quite useful for some analyses
  - well aligned with common practices of CERN experiments
- Solid developer base and community involvement

# Zenodo: should we use it?

- Please take a look at the system
- It's a policy decision - the capability is there
- Appears to have enough long-term viability to meet the DAP needs of PHENIX
- We don't lose control over original documents since the original files are still stored at BNL
- Appears to meet all reasonable requirements (versioning, tagging, extensive search capabilities, access control etc)
- A good pilot project would be to create a collection of PhD theses with proper tags
- Suggestions?

# Summary

- We (the DAP task force) have identified a suitable Web platform in which keep long-term, static documentation such as detector and run information, software manuals and other such materials. Participation of the PHENIX community in populating the new site with information will be crucial for its success. A collaborative environment (a repo on GitHub) is in place.
- There is a platform for managing documents which is readily available and feature-rich. It is hosted at CERN, is free and is expected to remain in operation for a long time. Using this platform (Zenodo) is a decision of policy rather than capability. Hosting PhD theses on Zenodo could be a pilot project and will create reliable references via DOI.
- Common features of the solutions above
  - they are state-of-the-art
  - they offer free and reliable hosting in foreseeable future.
- Both items have the potential to aid analyses in short and medium term as well as help PHENIX meet its long-term data preservation mandate.

# Backup slides

# PHENIX Analysis Notes Archive - the query page



## 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 publication

Use this form to search for technical and analysis notes.

Select first the desired note type either analysis notes or technical 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:

Search by:  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:

Collision Species/Energy:

Physics Working Group:

Analysis Type:

# PHENIX Analysis Notes - the archive contents

Number	Date	Title	Authors	Key Word	Links
<a href="#">an1425</a>	2019-10-08	Run15 pAu identified pion and anti-proton spectra	Weizhuang Peng, Julia Velkovska	PLHF, p+Al_200GeV, Run-15, identified particles	<a href="#">an1425</a> <a href="#">draft</a> <a href="#">an1282</a>
<a href="#">an1424</a>	2019-10-08	Run15 pAu identified pion and anti-proton spectra			<a href="#">an1424</a> <a href="#">draft</a>
<a href="#">an1423</a>	2019-10-08	Run15 pAu identified pion and anti-proton spectra			<a href="#">an1423</a> <a href="#">draft</a>
<a href="#">an1422</a>	2019-10-08	Run15 pAu identified pion and anti-proton spectra			<a href="#">an1422</a> <a href="#">draft</a>
<a href="#">an1421</a>	2019-10-08	Run15 pAu identified pion and anti-proton spectra			<a href="#">an1421</a> <a href="#">draft</a>
<a href="#">an1420</a>	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	<a href="#">an1420</a> <a href="#">draft</a>
<a href="#">an1419</a>	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 1pp 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	<a href="#">an1419.01</a> <a href="#">draft</a> <a href="#">an1152</a> <a href="#">an1269</a> <a href="#">an1270</a>
<a href="#">an1418</a>	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	<a href="#">an1418.02</a> <a href="#">draft</a>
<a href="#">an1417</a>	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	<a href="#">an1417.01</a> <a href="#">draft</a>
<a href="#">an1414</a>	2019-09-18	Template for PHENIX Analysis Notes	Yasuyuki Akiba, Gabor David		<a href="#">an1414</a> <a href="#">draft</a>
<a href="#">an1413</a>	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	<a href="#">an1413.01</a> <a href="#">draft</a>
<a href="#">an1412</a>	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	<a href="#">an1412</a> <a href="#">draft</a> <a href="#">an065</a> <a href="#">an1010</a> <a href="#">an1374</a> <a href="#">an1401</a> <a href="#">an1402</a> <a href="#">an770</a> <a href="#">an911</a> <a href="#">an964</a> <a href="#">PPG148</a>
<a href="#">an1411</a>	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	<a href="#">an1411</a> <a href="#">draft</a> <a href="#">an1074</a> <a href="#">an1231</a> <a href="#">an1260</a> <a href="#">an1374</a> <a href="#">an683</a> <a href="#">an814</a> <a href="#">PPG146</a>
<a href="#">an1410</a>	2019-10-24	Direct, elliptic and triangular flow of $\pi^0 \rightarrow \gamma\gamma$ 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	<a href="#">an1410.02</a> <a href="#">draft</a> <a href="#">an1367</a> <a href="#">an1406</a> <a href="#">an1407</a>
<a href="#">an1409</a>	2019-06-24	Run14 AuAu EMCal Geometry Tuning	Gabor David, Axel Drees, Wenqing Fan	PLHF, Au+Au_200GeV, Run-14, direct photons	<a href="#">an1409</a> <a href="#">draft</a>
<a href="#">an1408</a>	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	<a href="#">an1408.03</a> <a href="#">draft</a> <a href="#">an1273</a> <a href="#">PPG206</a> <a href="#">PPG221</a>
<a href="#">an1407</a>	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	<a href="#">an1407</a> <a href="#">draft</a> <a href="#">an1304</a>
<a href="#">an1406</a>	2019-05-27	Q vector calibration	Veronica CanoaRoman, Jaehyeon Do, Carlos PerezLara	Heavy Ion, d+Au_200GeV, Run-16, correlations	<a href="#">an1406</a> <a href="#">draft</a>
<a href="#">an1405</a>	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	<a href="#">an1405.04</a> <a href="#">draft</a>
<a href="#">an1404</a>	2019-05-08	Measurement and analysis of three-pion HBT correlations for 0-30% Centrality in 200 GeV AuAu collisions	Mate Csanad, Bálint Kurgys	PLHF, Au+Au_200GeV, Run-10, correlations	<a href="#">an1404.01</a> <a href="#">draft</a> <a href="#">an1187</a> <a href="#">an1244</a> <a href="#">an1288</a> <a href="#">an920</a> <a href="#">PPG194</a>
<a href="#">an1403</a>	2019-08-15	S/Jpsi as a function of $\sqrt{s_{NN}}$ in small systems with Yue Hang Leung's Correlated Background: Run15pp and Run15pAu, Run15pAl, Run14S*(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	<a href="#">an1403.06</a> <a href="#">draft</a> <a href="#">an1306</a> <a href="#">an1354</a> <a href="#">an1369</a> <a href="#">an1391</a>
<a href="#">an1402</a>	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	<a href="#">an1402.01</a> <a href="#">draft</a> <a href="#">an065</a> <a href="#">an1010</a> <a href="#">an1374</a> <a href="#">an770</a> <a href="#">an911</a> <a href="#">an964</a> <a href="#">PPG148</a>
				HHJ, PLHF, Au+Au_200GeV,	<a href="#">an1401.01</a> <a href="#">draft</a> <a href="#">an1174</a> <a href="#">an1308</a> <a href="#">an560</a>

# Invenio@BNL

- Test deployment of Invenio v3 at BNL by SDCC experts (C. Gamboa)
- An application developed by BNL SDCC for one of the BNL departments
  - some specific requirements to the document workflow, not necessarily applicable to other projects
  - work in progress now to make this work for sPHENIX (custom app), discussions ongoing
- Evaluation for the BNL EIC working group is yet to happen but is planned

# The role of BNL SDCC

- BNL SDCC will be (already effectively is) the “bit preservation” site for PHENIX
- In recent years BNL SDCC serves as the single fully functional center of the PHENIX computing infrastructure and provides critical services to this and other experiments. At present virtually all of the stages of data processing and analysis (including simulation) take place at BNL’s SDCC
- A variety of data products necessary for operation of PHENIX software (databases, files containing conditions/calibrations data etc) are stored and managed by SDCC
  - complete removal of dependencies would be very costly
- For that reason, any DAP plan in PHENIX needs to assume that future reuse and analysis of the data will take place at SDCC
- SDCC has a versatile team of experts that help solve problems outside of core production activities
  - a recent example is installation and evaluation of Invenio-based systems at BNL