

Zenodo: an update

Maxim Potekhin

BNL

Nuclear and Particle Physics Software Group

06/04/2020

The current situation

- An urgent (and long standing) need for a robust document repository in EIC community, in support of the Yellow Report effort and beyond
 - DocDB no longer an option, Zenodo (as previously reported) is the prime candidate for this role
 - CERN developed and based, a portable version (“RDM”) coming later this year
- Prompted by the need to support COVID-19 research at BNL the SDCC team has accomplished a custom installation of Zenodo at BNL, and just created a test instance for EIC
 - <https://eic-zenodo.sdcc.bnl.gov/> (BNL intranet)
 - Unclear why a Lab-wide DB is not sufficient, need to find out - it’s content-agnostic
 - Zenodo was not designed for portability so the install is custom
 - Integration with local auth/auth, SSO etc - may be helpful, may be not
 - “Invenio RDM” which is the next generation of this system will be truly portable but ETA is late 2020 which is late for the Yellow Report and other immediate purposes
 - Migration to RDM should work since CERN will need to do it first for their instance

First experience with Zenodo

- Meets the definition of a good system:
 - Simple things are “easy”
 - Complex things are “possible”
- Initial learning curve is quick and painless, the keyword function transparent while additional extensive search capabilities are available
- Versioning, ease of metadata editing
- ORCID and DOI capability off the bat (doi.org)
- Tiers of access
 - Private (locked in)
 - Restricted (by individual request)
 - Embargo (time out of restrictions)
 - Public
- “Communities” - groups of documents curated by designated persons
- GitHub integration - citeable code with permanent DOI and link

Zenodo is not a document workflow system

- It is a repository, not a document workflow management system
- i.e. there is no document development pipeline with comments, approvals and tiered access to specific versions
- I consider it highly unlikely that such functionality will ever be part of Zenodo because it's not aligned with its mission
 - What is published is expected to be close to a finished product, which can be further refined by cutting versions
- As an aside comment, much of the document workflow management can be already achieved with the issue tracking system on GitHub, perhaps combined with a private repository and controlled access
 - State-of-the-art, robust system
 - Comments, replies, commits

On Zenodo, you cannot permanently delete a record

- Since the DOI capability (conceptually permalinks) is an integral part of the system the material is not allowed to completely vanish
- ...but can be made permanently unavailable for viewing by anyone
- This means some thought needs to be put into what is committed to Zenodo just to avoid clutter going forward
- Robust search capabilities also mean that dummy/dark records are easy to avoid in practice

DOI, versions, keywords, conference-awareness

March 2, 2020

Evolution of the Data Quality Monitoring and Prompt Processing System in the protoDUNE-SP experiment

Maxim Potekhina

The DUNE Collaboration has successfully implemented and currently operates an experimental program based at CERN which includes a beam test and an extended cosmic ray run of two large-scale prototypes of the DUNE Far Detector. The volume of data already collected by the protoDUNE-SP (the single-phase Liquid Argon TPC prototype) amounts to approximately 3PB and the sustained rate of data sent to mass storage is of the order of 0(100) MB/s. In addition to this data being committed to mass storage and processed in the Grid environment a small fraction of it is captured by the Prompt Processing System which is optimized for continuous low-latency calculation of the vital detector metrics and parameters as well as the output rendered as event display images. This system is the platform for Data Quality Monitoring in protoDUNE-SP and has served a crucial role starting from the commissioning of the apparatus and throughout its operation in 2018-2019, which continues at the time of writing. We present our experience in operating the system in the CERN environment, as well as work currently underway to make the system more scalable, resilient and to simplify system recovery procedures in preparation for the second run of protoDUNE-SP foreseen after the Long Shutdown of the LHC in the Fall of 2019.

18 views 11 downloads
[See more details...](#)

Indexed in
OpenAIRE

Publication date:
March 2, 2020

DOI:
DOI 10.5281/zenodo.3693788

Keyword(s):
dqm prompt processing data quality monitoring DUNE neutrino Liquid Argon

Meeting:
24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019), Adelaide, Australia, 4-8 November 2019

License (for files):
Creative Commons Attribution 4.0 International

Versions
Version 1 Mar 2, 2020
10.5281/zenodo.3693788

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.3693787. This DOI represents all versions, and will always resolve to the latest one. Read more.

Keywords

Edit/New Version

DOI

Conference

Version

Preview

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:

- 3D event display based on raw data
- The TPC monitoring application with an extensive set of histograms at various levels of channel aggregation, FFT charts etc
- Health monitor for the Front End Electronics monitoring cards
- Prompt reconstruction with metrics such as track candidate count, hit count etc
- Track candidate based Argon purity estimations with time series stored in the DB
- Signal-to-noise ratio monitoring
- Data preparation for the 3D event display implemented on a separate server
- Out-of-band feed from the hardware Argon purity monitor

The DQM application platform (LArSoft) was periodically updated and the software provisioning was done via CVMFS. Additions and modifications of the DQM payloads were facilitated by metadata generated by applications which was used to automatically generate menus and layouts for graphics on Web pages, without changes on the server side.

Evolving monitoring

Monitoring tools were added to the prompt processing system to address a few infrastructure issues encountered during its operation: e.g. selecting metadata for file transfer (alioth), batch system performance, storage required by the input data and displaying alarms to operators as necessary.

3D event display

Health monitor for FEET

Signal-to-noise ratio

Track candidate based Argon purity

OpenDash Dashboard

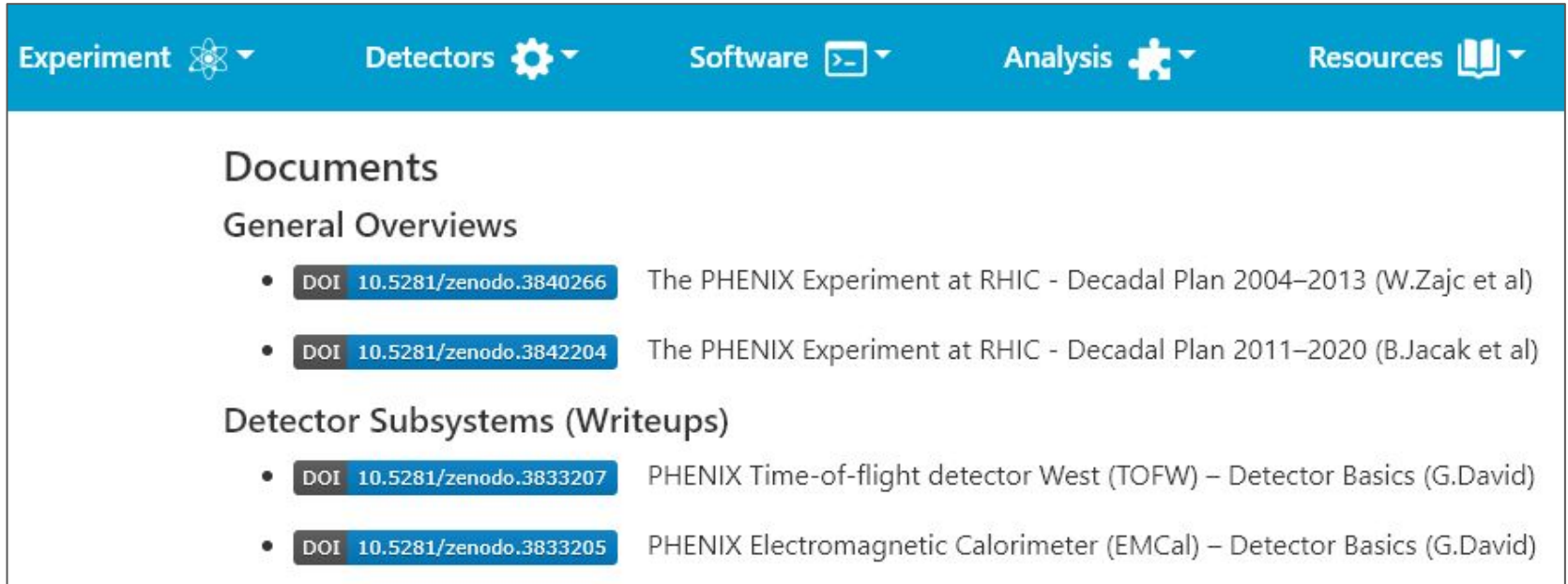
FFT

Recent Zenodo activities in PHENIX






- Approved for PHENIX Data and Analysis Preservation (EC, conveners)
 - Using the CERN instance, location of the host not considered an issue
 - Good discoverability/visibility
- “Zenodo Communities”
 - A “PHENIX Collaboration” community created, managed by the team
 - DAP site links to Zenodo (next slide)
 - In general, a nice way to link to materials from websites
- GitHub integration - “nice to have” but not core - initial testing done
 - Additional cloud replica of your GitHub release tagged with arbitrary metadata (discoverability)
 - Citeable via DOI

Zenodo integration in with the PHENIX DAP website

- PHENIX is using Zenodo for **Data and Analysis Preservation**. We find that it's simple to use and meets most of the needs of Collaboration.
- Created simple tools to easily reference Zenodo records on PHENIX sites



The screenshot displays the PHENIX DAP website interface. At the top is a blue navigation bar with five menu items: 'Experiment' (with a particle icon), 'Detectors' (with a gear icon), 'Software' (with a terminal icon), 'Analysis' (with a puzzle piece icon), and 'Resources' (with a book icon). Below the navigation bar, the 'Documents' section is visible, containing two sub-sections: 'General Overviews' and 'Detector Subsystems (Writeups)'. Each sub-section lists documents with their DOI and a brief description.

Experiment  **Detectors**  **Software**  **Analysis**  **Resources** 

Documents

General Overviews

- DOI [10.5281/zenodo.3840266](https://doi.org/10.5281/zenodo.3840266) The PHENIX Experiment at RHIC - Decadal Plan 2004–2013 (W.Zajc et al)
- DOI [10.5281/zenodo.3842204](https://doi.org/10.5281/zenodo.3842204) The PHENIX Experiment at RHIC - Decadal Plan 2011–2020 (B.Jacak et al)


Detector Subsystems (Writeups)


- DOI [10.5281/zenodo.3833207](https://doi.org/10.5281/zenodo.3833207) PHENIX Time-of-flight detector West (TOFW) – Detector Basics (G.David)
- DOI [10.5281/zenodo.3833205](https://doi.org/10.5281/zenodo.3833205) PHENIX Electromagnetic Calorimeter (EMCal) – Detector Basics (G.David)

Zenodo Community (another way to tag material)

- A way to organize material, and to consistently attribute materials to a collaboration/project/experiment - keeping a consistent brand
 - No need for multiple Zenodo instances?
- An improvement in visibility/discoverability/PR
 - An addition to the already existing metadata query aids in discovery of materials
- Anyone can upload a material to the community which is subject to **curation**
 - The curator gets notified and inspects the submission
 - If accepted, it becomes posted under the community umbrella
 - If rejected, it still remains on Zenodo site but is not officially owned/acknowledged by the community, this is an accordance to the “open access” platform
 - *There is currently one curator per community and there is no easy way to transfer this duty to a different account* (something few people expected) but a fix is on the way according to the lead developer and other team members. Unofficial ETA is late 2020.

PHENIX Community on Zenodo

[Upload](#)[Communities](#)

 potekhin@bnl.gov

The PHENIX Collaboration Community

Recent uploads

[View](#)

April 21, 2020 (v1.0) [Software](#) [Open Access](#)

PhenixCollaboration/web: First release of the PHENIX DAP site
Maxim Potekhin; Ron Belmont; amolhj
This is the first release of the PHENIX DAP website
Uploaded on April 21, 2020

[View](#)

May 1, 2019 (v1) [Thesis](#) [Open Access](#)

Transverse, Single-Spin Asymmetries for Charged Hadrons and for Muons from Open-Heavy-Flavor Decays in Polarized Proton-Proton and Proton-Nucleus Collisions in PHENIX
Bok, Jeongsu.
Transverse single-spin asymmetry (TSSA) phenomena have gained substantial attention in several decades because they provide valuable information on the spin structure of the nucleon. Production of heavy flavor is dominated by gluon-gluon fusion in the leading order perturbative Quantum Chromodynamic
Uploaded on April 20, 2020

[View](#)

May 1, 2019 (v1) [Thesis](#) [Open Access](#)


Measurements of $\mu\mu$ pairs from cc , $b\bar{b}$ and Drell-Yan in p+p and p+Au collisions at $\sqrt{s_{NN}} = 200$ GeV with PHENIX at RHIC
Leung, Yue Hang.
Dilepton spectra are a classic probe to study ultra-relativistic heavy ion collisions. At $\sqrt{s_{NN}} = 200$ GeV, the dilepton continuum is dominated by correlated pairs from semi-leptonic decays of charm and bottom hadrons and the Drell-Yan process. Measuring the azimuthal correlations of heavy flavor
Uploaded on April 20, 2020

[View](#)

[More](#)

[New upload](#)

Community



The PHENIX Collaboration Community

The PHENIX Collaboration has initiated a Data and Analysis Preservation (DAP) effort in 2019. Within this scope there are a few areas of activity such as curating the available information, development of a new website to support DAP and to systematize and store available document for the long term. The latter is the main reason the PHENIX Zenodo community has been created.

Curated by:
MaximPotekhin

Curation policy:
Curation is done by members of the PHENIX DAP Task Force and contributors who join individual projects within the scope of DAP.

Created:
April 20, 2020

Harvesting API:
[GAI-PMH Interface](#)

Want your upload to appear in this community?

[Click the button above to upload straight to](#)

Advanced search capabilities

by default, all searches are sorted according to an internal ranking algorithm that scores each match against your query. In both the user interface and REST API, it's possible to sort the results by:

- Most recent
- Publication date
- Title
- Conference session
- Journal
- Version

Regular expressions

Regular expressions are a powerful pattern matching language that allow to search for specific patterns in a field. For instance if we wanted to find all records with a DOI-prefix 10.5281 we could use a regular expression search:

Example: `doi:10\..5281V.+/`

Careful, the regular expression must match the *entire* field value. See the [regular expression syntax](#) for further details.

Missing values

It is possible to search for records that either are missing a value or have a value in a specific field using the `_exists_` and `_missing_` field names.

Example: `_missing_.notes` (all records without notes)

Example: `_exists_.notes` (all records with notes)

Advanced concepts

Boosting

You can use the boost operator `^` when one term is more relevant than another. For instance, you can search for all records with the phrase *open science* in either *title* or *description* field, but rank records with the phrase in the *title* field higher:

Example: `title:"open science"~5 description:"open science"`

Fuzziness

You can search for terms similar to but not exactly like your search term using the fuzzy operator `~`.

Example: `oepr~`

Results will match records with terms similar to `oepr` which would e.g. also match `open`.

Proximity searches

A phrase search like `"open science"` by default expect all terms in exactly the same order, and thus for instance would not match a record containing the phrase *open access and science*. A proximity search allows that the terms are not in the exact order and may include other terms inbetween. The degree of flexibility is specified by an integer afterwards:

Example: `"open science"~5`

Wildcards

You can use wildcards in search terms to replace a single character (using `?` operator) or zero or more characters (using `*` operator).

Example: `ope? scien*`

Wildcard searches can be slow and should normally be avoided if possible.

Fields reference

The table below lists the data type of each field. Below is a quick description of what each data type means and what is possible.

- **string** Field does not require exact match (example field: `title`).

GitHub/Zenodo mechanics (see backup slides)

- A snapshot of a GitHub repo can be included in Zenodo organically+DOI
 - Integration/app link is in place: prepares and preserves tarballs of your releases
 - Makes your code easy to find (using the metadata) and to reference by a unique ID
 - Nice GUI
 - DOI reference to the code - becomes citeable
- Easy to use
 - Well-developed interface, I tested this functionality and it was quite simple
 - DOIs take some time $O(10\text{min})$ to propagate to the DOI.org system, but this is not a problem

Summary

- We need to be aware of Zenodo strengths and limitations
- Document workflow will need to be managed outside of Zenodo (I propose GitHub, potentially with private repos)
- Zenodo is a well designed system currently based at CERN with apparently good support and a long projected lifetime
 - “Communities” are a good way to curate and organize data
 - Lots of storage available
 - Any sort of data can be committed to Zenodo, with generous limits per upload
 - Multiple files in a single record
- A portable “RDM” release coming later this year
- A custom install at BNL available for test purposes
 - How is this all aligned with the Yellow Report process?



Terminology

- **Zenodo** is an open science data repository at CERN
 - In a nutshell, storage+metadata
 - Any data within the set limits
- **Invenio** is a toolkit used to in a number of CERN systems *including* Zenodo
 - A complex and capable framework.
 - Framework, not a system. *An application is needed to make use of its functionality.*
 - *cf. Zenodo is an Invenio-based application.*
- **Invenio RDM** (“research data management”) is a new product aiming to achieve
 - Portability (currently installing and configuring Invenio requires a high level of expertise)
 - Configurability i.e. eliminating the need for a custom app - a turnkey solution
 - ETA: late 2020

Zenodo “in a nutshell”

- General purpose digital repository
- Version control
- Data (storage space) + Metadata (DB)
- Extensive query capabilities
 - Full-text search is in the works
- DOI management (**doi.org** integration)
- ORCID-aware
- Gateway to other repositories
- GitHub integration (citeable code)
- Currently a service instance at CERN, being transformed into a more portable system under the “Invenio RDM” brand

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: durability

Safe

— more than just a drop box!

Your research output is stored safely for the future in same cloud infrastructure as research data from CERN's [Large Hadron Collider](#) and using CERN's battle-tested repository software [Invenio](#), which is used by some of the world's largest repositories such as [INSPIRE HEP](#) and [CERN Document Server](#).

Zenodo - GitHub panel - repo selection

The screenshot displays the Zenodo GitHub panel interface. At the top, the Zenodo logo is on the left, followed by a search bar and links for 'Upload' and 'Communities'. The user's email, 'potekhin@bnl.gov', is shown in the top right. Below the header, a breadcrumb trail reads 'Home / Account / GitHub'. On the left, a 'Settings' sidebar lists options: Profile, Change password, Security, Linked accounts, Applications, Shared links, and GitHub (which is selected). The main content area is titled 'GitHub Repositories' and includes a '(updated now)' status and a 'Sync now ...' button. A 'Get started' section with a circular arrow icon contains three steps: 1. Flip the switch (with an 'ON' button), 2. Create a release (with a link to 'create a release'), and 3. Get the badge (with an example DOI: 10.5281/zenodo.8475). Below this, a 'Repositories' section lists six repositories with their corresponding GitHub links and 'OFF' toggle switches: BNLNPPS/BNLNPPS.github.io, BNLNPPS/BirdView, BNLNPPS/tpc-rs, DUNE/FNALCore, DUNE/Sandbox-TDR, and DUNE/SpaceCharge.

zenodo Search Upload Communities potekhin@bnl.gov

Home / Account / GitHub

Settings

- Profile
- Change password
- Security
- Linked accounts
- Applications
- Shared links
- GitHub**

GitHub Repositories (updated now) Sync now ...

Get started


- 1 Flip the switch**
Select the repository you want to preserve, and toggle the switch below to turn on automatic preservation of your software.
☒ ON
- 2 Create a release**
Go to GitHub and [create a release](#). Zenodo will automatically download a zip-ball of each new release and register a DOI.
- 3 Get the badge**
After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.
DOI [10.5281/zenodo.8475](#) (example)


Repositories

If your organization's repositories do not show up in the list, please ensure you have enabled third-party access to the Zenodo application. Private repositories are not supported.



| | |
|---|------------------------------|
| BNLNPPS/BNLNPPS.github.io | <input type="checkbox"/> OFF |
| BNLNPPS/BirdView | <input type="checkbox"/> OFF |
| BNLNPPS/tpc-rs | <input type="checkbox"/> OFF |
| DUNE/FNALCore | <input type="checkbox"/> OFF |
| DUNE/Sandbox-TDR | <input type="checkbox"/> OFF |
| DUNE/SpaceCharge | <input type="checkbox"/> OFF |

Zenodo - GitHub panel - published release







[Upload](#) [Communities](#)


 potekhin@bnl.gov 


[Home](#) / [Account](#) / [GitHub](#) / [Repository](#)


Settings


 [Profile](#)


 [Change password](#)


 [Security](#)

 [Linked accounts](#)

 [Applications](#)


 [Shared links](#)


 **GitHub**


 **PhenixCollaboration/web**


DOI [10.5281/zenodo.3759876](#)


GitHub / Releases

 [Create release ...](#)

 **v1.0 PhenixCollaboration/web: First release of the PHENIX DAP site**

 **Published**

 DOI: [10.5281/zenodo.3759876](#)

 First release of the PHENIX DAP site

11 minutes ago

Zenodo - GitHub panel - published release browser

The screenshot shows the Zenodo interface for a specific release. The header is blue with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. Below the header, the date 'April 21, 2020' is displayed on the left, and 'Software' and 'Open Access' badges are on the right. The main title is 'PhenixCollaboration/web: First release of the PHENIX DAP site', followed by the authors 'Maxim Potekhin; Ron Belmont; amolhj'. A description states 'This is the first release of the PHENIX DAP website'. A 'Preview' section is expanded, showing a file tree for 'web-v1.0.zip'. The tree lists various files and folders with their respective sizes.

zenodo Search Upload Communities

April 21, 2020 Software Open Access

PhenixCollaboration/web: First release of the PHENIX DAP site

Maxim Potekhin; Ron Belmont; amolhj

This is the first release of the PHENIX DAP website

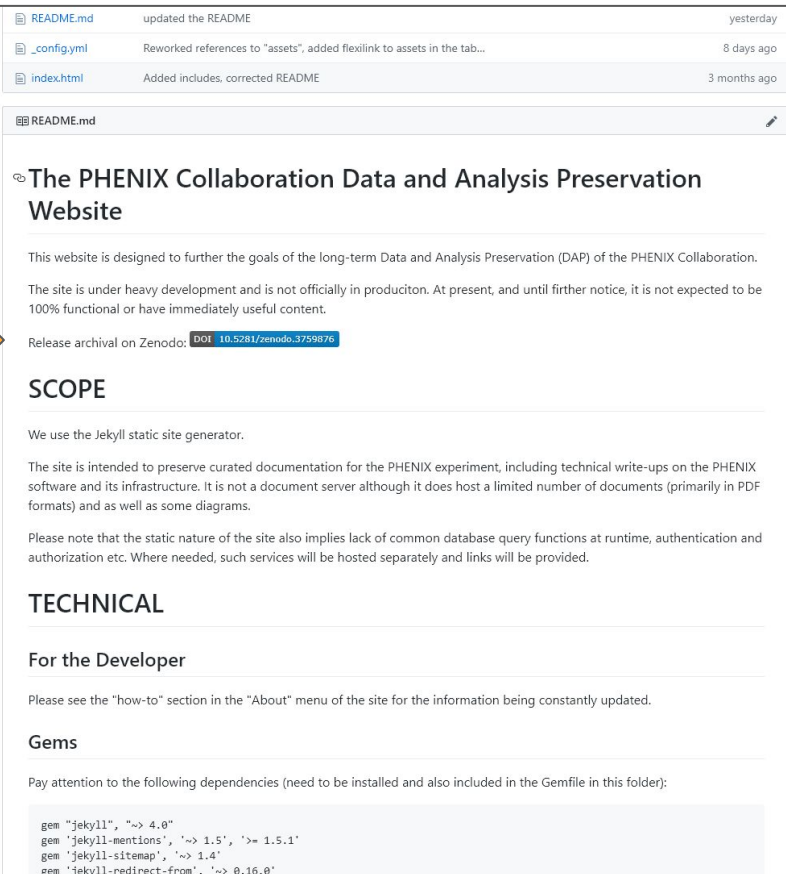
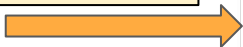
Preview

web-v1.0.zip

- PhenixCollaboration-web-c9d991e
 - .gitignore 216 Bytes
 - Gemfile 285 Bytes
 - LICENSE 11.4 kB
 - README.md 1.6 kB
 - _about
 - contact.md 324 Bytes
 - dap.md 1.3 kB
 - howto.md 6.9 kB
 - site.md 2.5 kB
 - _analysis
 - overview.md 114 Bytes
 - _config.yml 1.3 kB
 - _data
 - acc
 - vars.yml 1.4 kB
 - detectors.yml 2.6 kB
 - documents.yml 3.5 kB
 - gallery.yml 4.5 kB

DOIs are an increasingly popular way to reference software

Persistent, durable link to archived software, can be nicely embedded in any page.



The screenshot shows a GitHub repository page. At the top, there's a file list with three files: `README.md` (updated yesterday), `_config.yml` (updated 8 days ago), and `index.html` (updated 3 months ago). Below this is the `README.md` file content. The title is "The PHENIX Collaboration Data and Analysis Preservation Website". The text describes the website's purpose for long-term data and analysis preservation. A key feature is the release archival on Zenodo, with a DOI link: [DOI 10.5281/zenodo.3759876](https://doi.org/10.5281/zenodo.3759876). The page is divided into sections: "SCOPE", "TECHNICAL", "For the Developer", and "Gems". The "Gems" section lists dependencies for the Jekyll static site generator.

```
gem "jekyll", "~> 4.0"
gem "jekyll-mentions", "~> 1.5", ">= 1.5.1"
gem "jekyll-sitemap", "~> 1.4"
gem "jekyll-redirect-from", "~> 0.16.0"
```

GitHub/Zenodo integration benefits

- Not a core functionality by a long shot, however...
- ...provides a uniform way to reference digital products using DOI
- ...metadata is a good thing to have - better discoverability!
- ...can leverage the Zenodo “community” feature to organize materials and increase visibility
 - Cf. simulated data and the code used to produce it can be kept under the same umbrella
- Longer term - Data and Analysis Preservation
- In general, an “EIC Software” community on Zenodo may be a useful thing to have (papers, conference presentations etc)