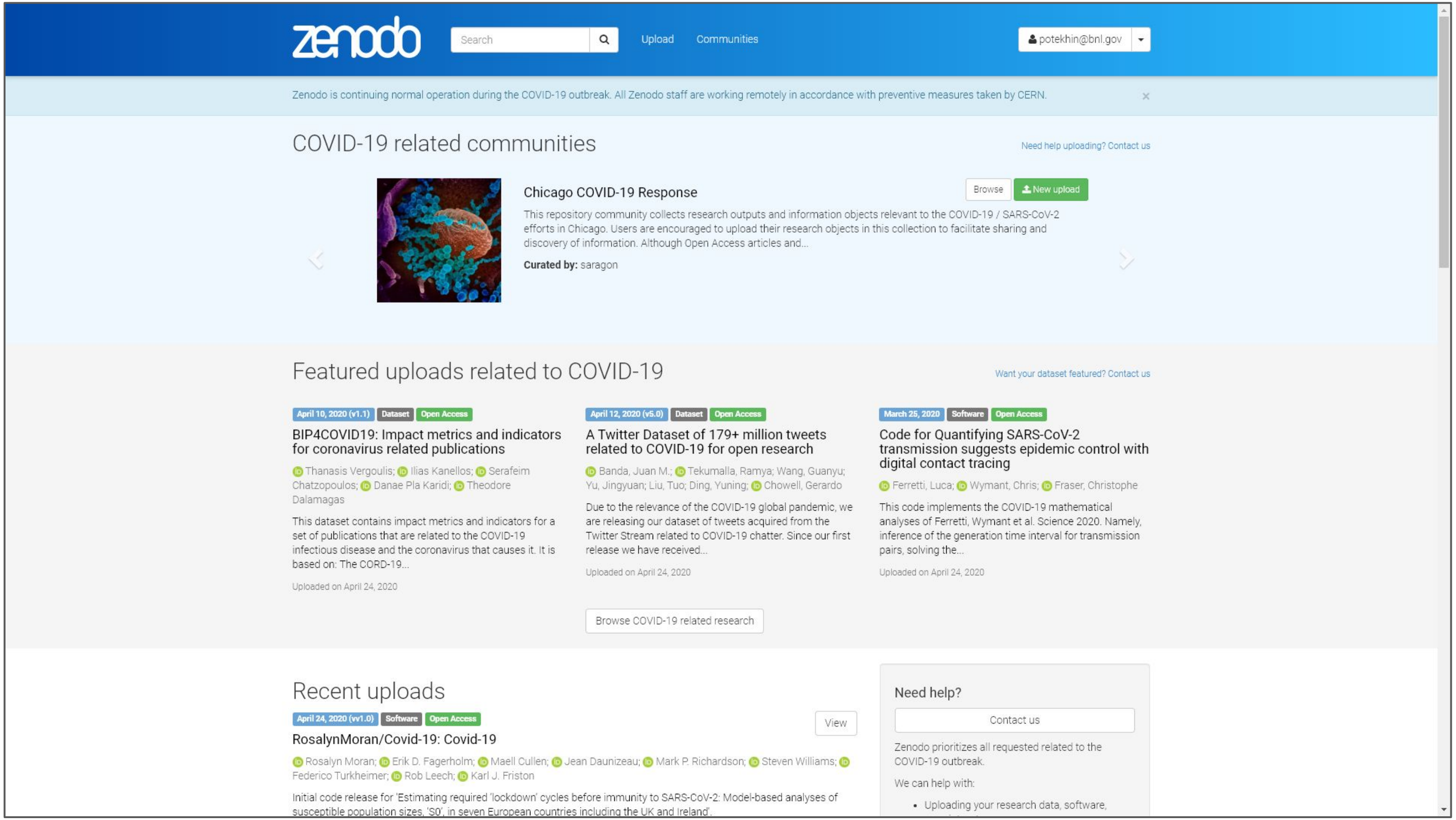
Zenodo/Invenio for EIC: Yellow Report and beyond

Maxim Potekhin
(BNL, NPPS)
04/29/2020

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

https://zenodo.org/ - named after Ζηνόδοτος, inventor of metadata in 280 BC



Zenodo is continuing normal operation during the COVID-19 outbreak. All Zenodo staff are working remotely in accordance with preventive measures taken by CERN.

COVID-19 related communities

Need help uploading? Contact us

Chicago COVID-19 Response

This repository community collects research outputs and information objects relevant to the COVID-19 / SARS-CoV-2 efforts in Chicago. Users are encouraged to upload their research objects in this collection to facilitate sharing and discovery of information. Although Open Access articles and...

Curated by: saragon

Featured uploads related to COVID-19

Want your dataset featured? Contact us

April 10, 2020 (v1.1) Dataset Open Access

BIP4COVID19: Impact metrics and indicators for coronavirus related publications

Thanasis Vergoulis; Ilias Kanellos; Serafeim Chatzopoulos; Danae Pla Karidi; Theodore Dalamagas

This dataset contains impact metrics and indicators for a set of publications that are related to the COVID-19 infectious disease and the coronavirus that causes it. It is based on: The CORD-19...

Uploaded on April 24, 2020

April 12, 2020 (v5.0) Dataset Open Access

A Twitter Dataset of 179+ million tweets related to COVID-19 for open research

Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Chowell, Gerardo

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received...

Uploaded on April 24, 2020

March 25, 2020 Software Open Access

Code for Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing

Ferretti, Luca; Wymant, Chris; Fraser, Christophe

This code implements the COVID-19 mathematical analyses of Ferretti, Wymant et al. Science 2020. Namely, inference of the generation time interval for transmission pairs, solving the...

Uploaded on April 24, 2020

Recent uploads

April 24, 2020 (vv1.0) Software Open Access

RosalynMoran/Covid-19: Covid-19

Rosalyn Moran; Erik D. Fagerholm; Maell Cullen; Jean Daunizeau; Mark P. Richardson; Steven Williams; Federico Turkheimer; Rob Leech; Karl J. Friston

Initial code release for 'Estimating required 'lockdown' cycles before immunity to SARS-CoV-2: Model-based analyses of susceptible population sizes, 'S0', in seven European countries including the UK and Ireland'.

Need help?

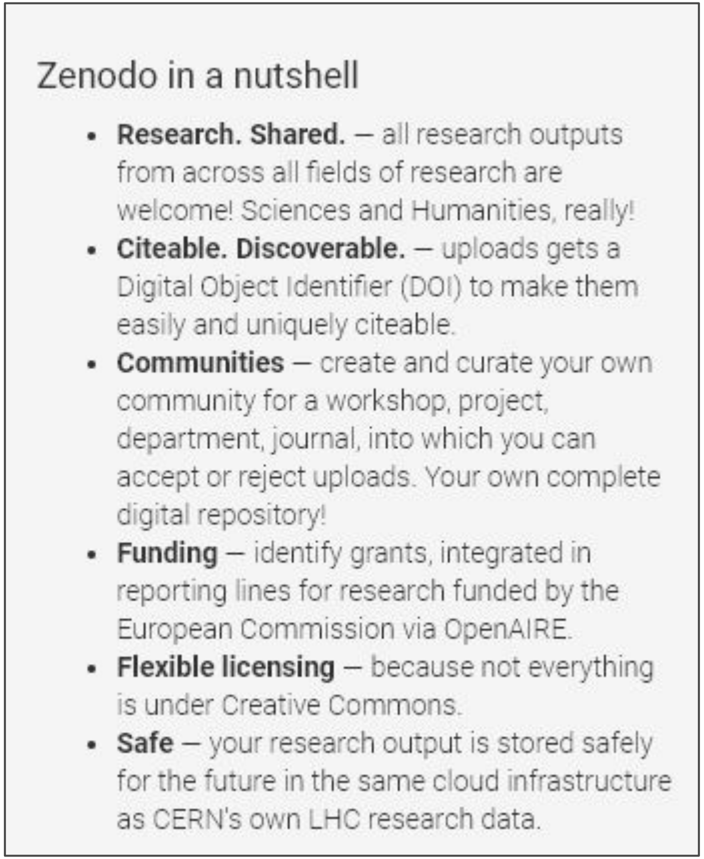
Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

- Uploading your research data, software,

Zenodo “in a nutshell”

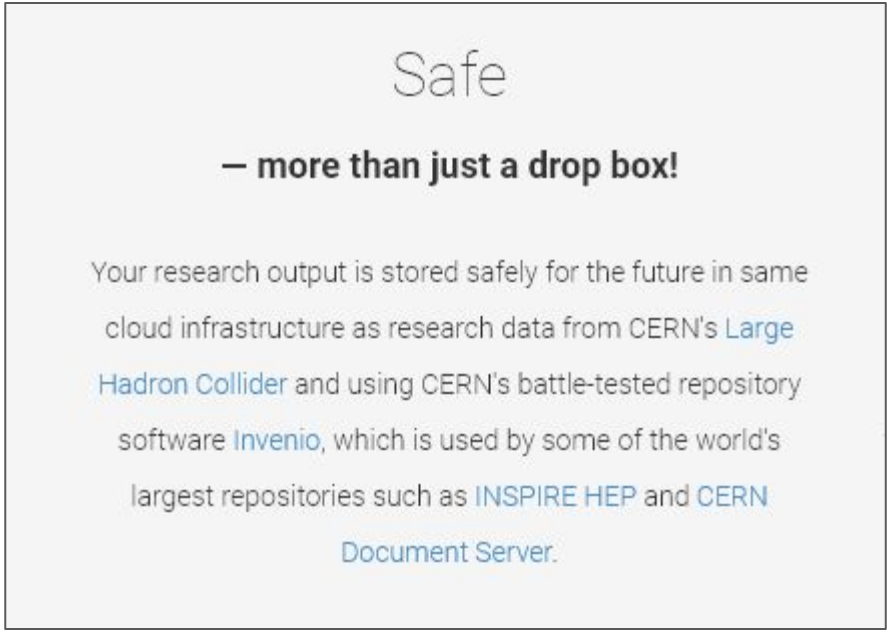
- General purpose digital repository
- Version control
- Data (storage space) + Metadata (DB)
- Extensive query capabilities
  - Full-text search is in the works
- DOI management (**doi.org** integration)
- ORCID-aware
- Gateway to other repositories
- GitHub integration (citeable code)
- Currently a 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

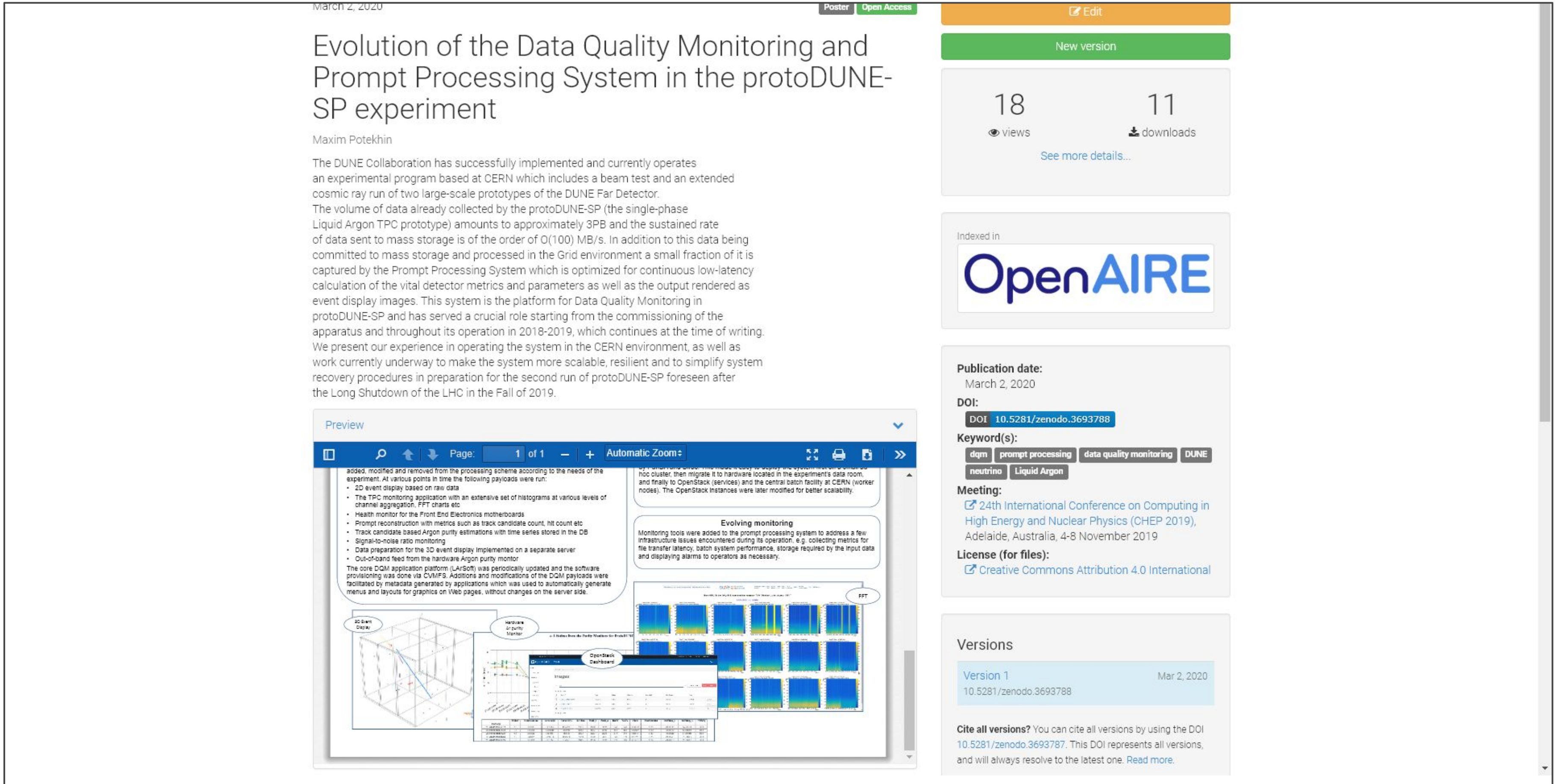


Safe

— 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.

DOI, keywords, conference-awareness



March 2, 2020

Poster Open Access

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

Maxim Potekhin

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 O(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

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:

- 2D 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 motherboards
- 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 montor

The core 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.

hoc cluster, then migrate it to hardware located in the experiment's data room, and finally to OpenStack (services) and the central batch facility at CERN (worker nodes). The OpenStack instances were later modified for better scalability.

Evolving monitoring

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

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.

Motivations

- Managing documents and other materials is a universal necessity in the field
  - Consider the needs of the Yellow Report working groups (papers, presentations, tables etc)
  - Not a replacement of the Wiki (which is not a document handling system in the first place)
- Not too many products exist in that area
  - DocDB is used at FNAL, BNL and a few other places, it’s an aging product, no clear API
  - CERN CDS is not portable (NB shares the Invenio back-end with Zenodo)
- In EICUG here is currently not a single accepted solution or a policy
- The new EIC Software website is not designed as a general purpose document store (scalability, lack of proper metadata etc)
- Zenodo is an obvious contender

Invenio at BNL

- BNL SDCC is testing installation of **Invenio RDM**, BNL being an official partner as a test site
  - Install is currently pretty hard and not ready for prime time
  - In fairness, this is a complex product with many moving parts
- sPHENIX moved to use a custom app based on pre-RDM Invenio to use in lieu of DocDB
- In more recent news, Zenodo instance(s) have been created
  - TBD in an upcoming meeting

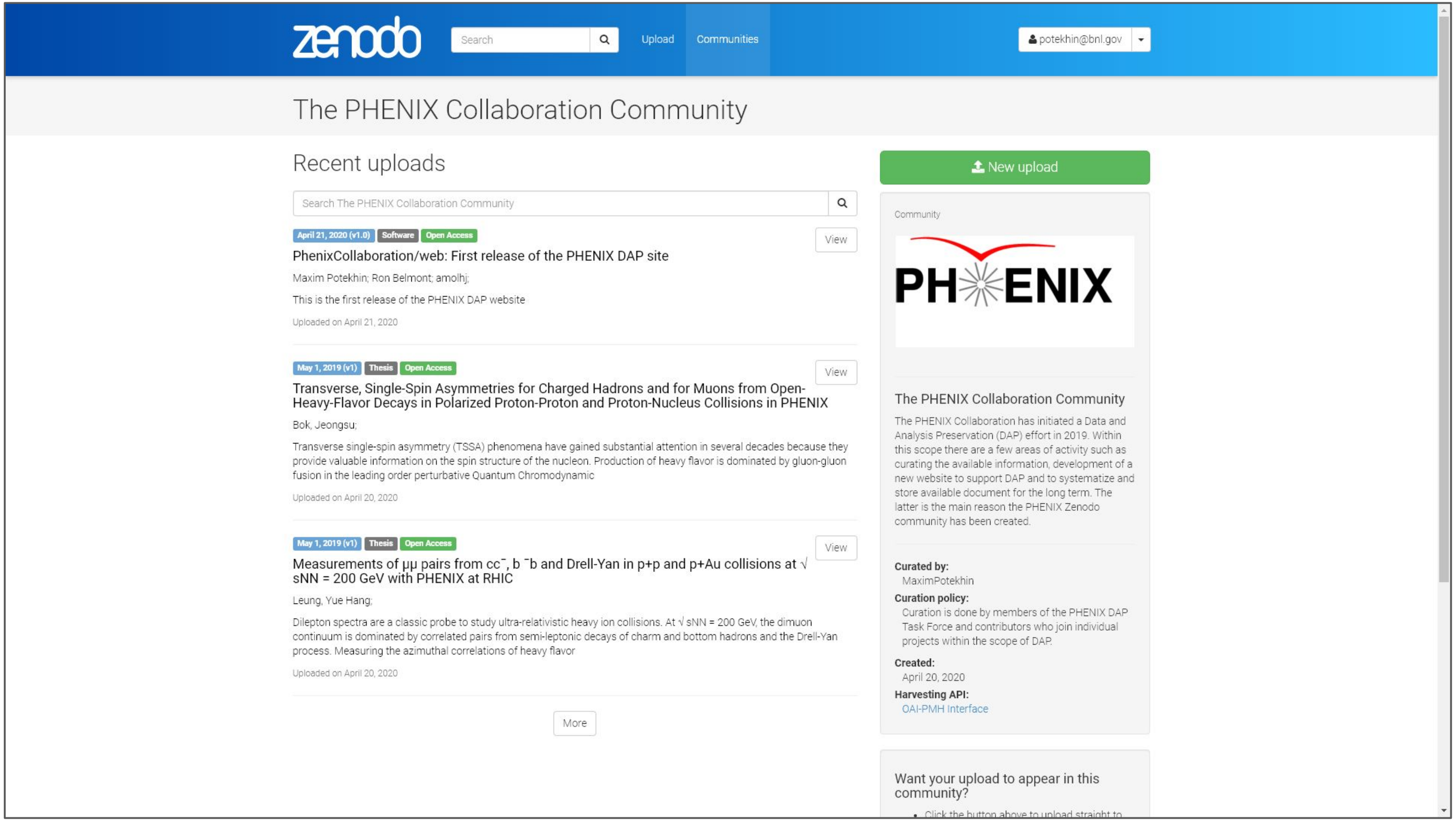
Recent Zenodo activities

- Approved for PHENIX Data and Analysis Preservation (DAP)
- “Zenodo Communities” - see next slide - functional testing started
    - A “PHENIX Collaboration” community created, started populating it with materials
- Communication with the developers, looking for guidance regarding
    - Possible future data migration from Zenodo to Invenio RDM
    - Feature requests for community management
    - Storage allocation and use pattern discussion
- 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 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
- 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.

A Community Example



The PHENIX Collaboration Community

Recent uploads

Search The PHENIX Collaboration Community

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

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

May 1, 2019 (v1) | Thesis | Open Access

Measurements of μμ pairs from cc¯, b ¯b and Drell-Yan in p+p and p+Au collisions at √ sNN = 200 GeV with PHENIX at RHIC

Leung, Yue Hang;

Dilepton spectra are a classic probe to study ultra-relativistic heavy ion collisions. At √ sNN = 200 GeV, the dimuon 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

Community

PHENIX

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.

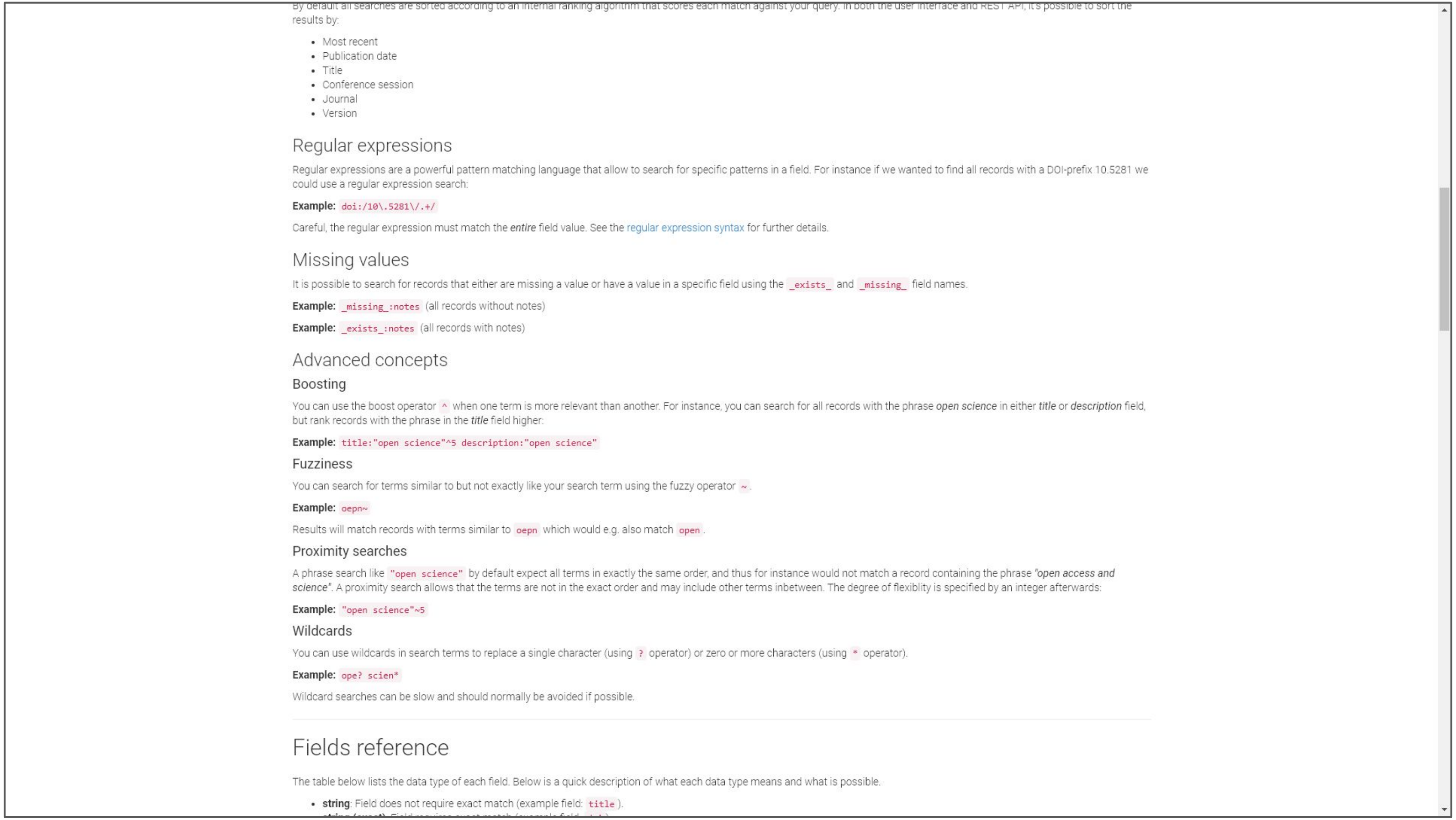
Created:
April 20, 2020

Harvesting API:
OAI-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\.5281\/.+/`

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: `oepn~`

Results will match records with terms similar to `oepn` 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 flexiblity 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`).
- **string (exact)**: Field requires exact match (example field:

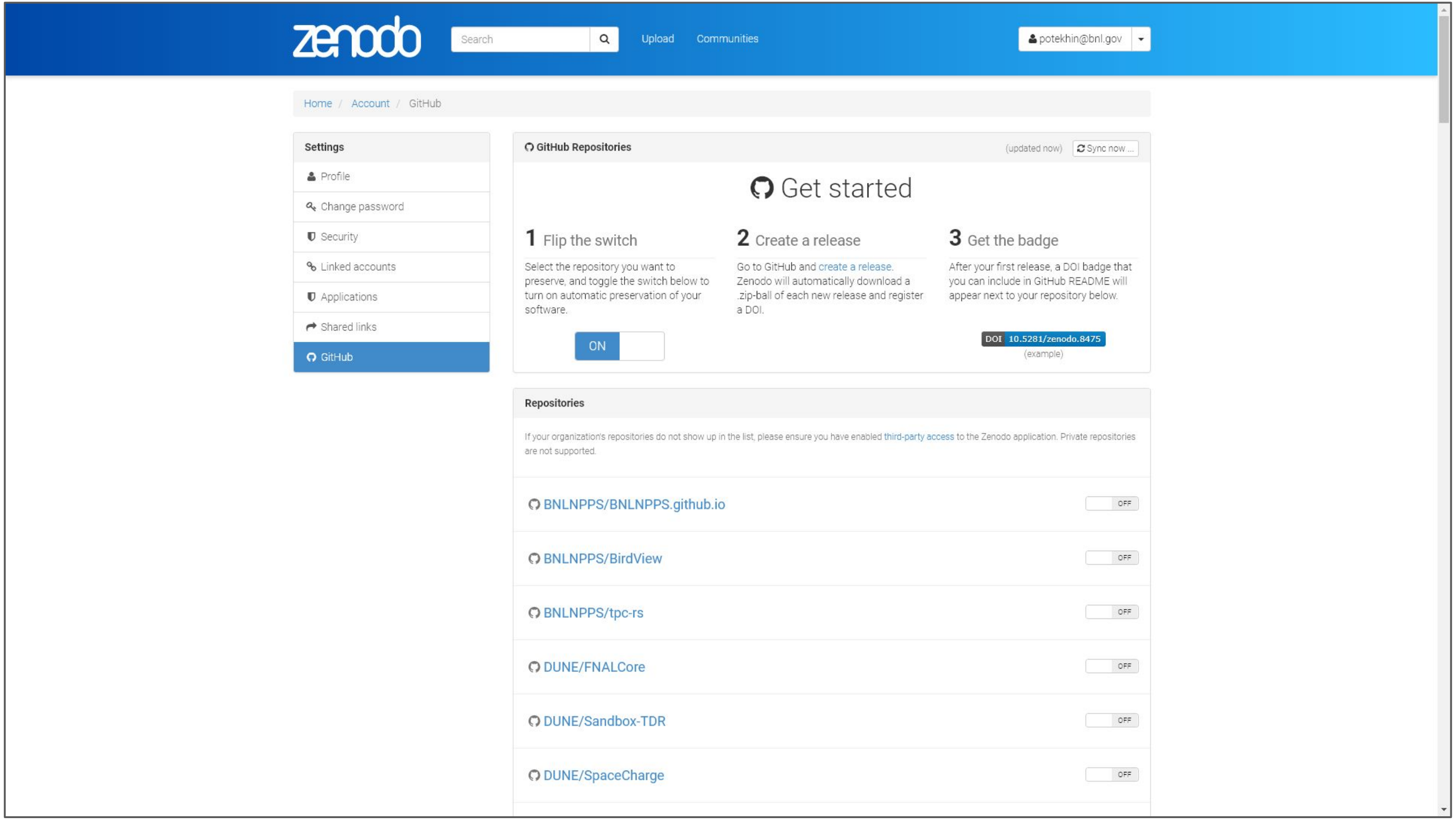
Policy issues

- Zenodo defines itself as an open science platform i.e. for the most part public
- It does have access tiers: private, restricted and public
  - “Restricted” means that a request for access is forwarded to the owner
  - Not designed to handle “roles” for large groups of people
- Consider the fact that DocDB instances are often protected
  - In reality I would say 95%+ of materials don’t need to be protected

GitHub/Zenodo mechanics

- A snapshot of a GitHub repository 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
- Easy to use
  - I tested this functionality and it was quite simple
  - DOIs take some time O(10min) to propagate to the DOI.org system

Zenodo - GitHub panel - repo selection



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.

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

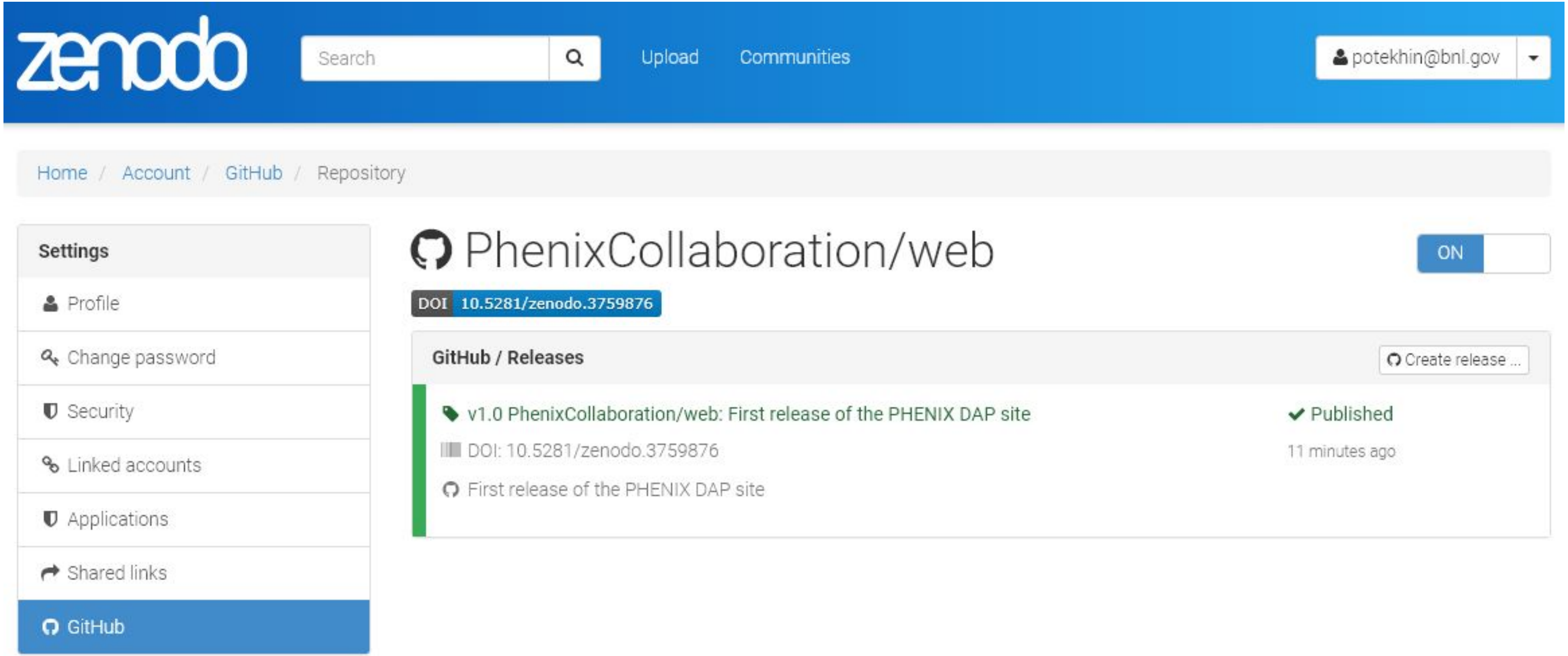
After your first release, a DOI badge that you can include in GitHub README will appear next to your repository below.

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 — OFF
- BNLNPPS/BirdView — OFF
- BNLNPPS/tpc-rs — OFF
- DUNE/FNALCore — OFF
- DUNE/Sandbox-TDR — OFF
- DUNE/SpaceCharge — OFF

Zenodo - GitHub panel - published release



PhenixCollaboration/web

DOI 10.5281/zenodo.3759876

GitHub / Releases

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

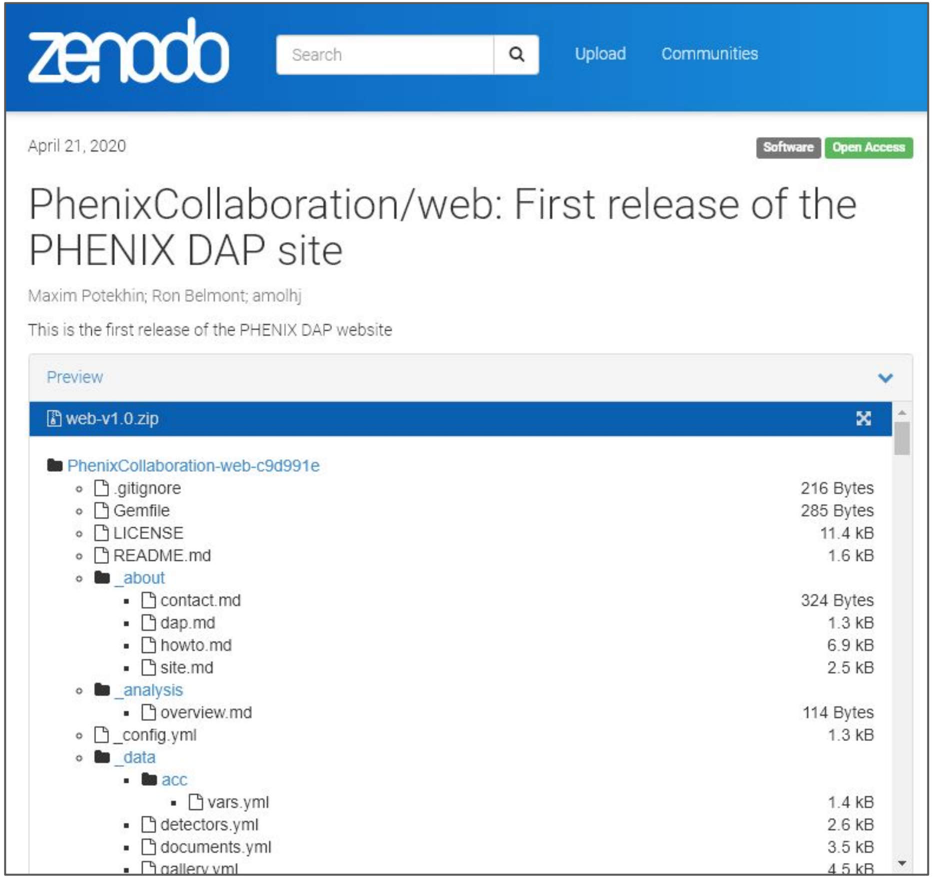
✓ Published

DOI: 10.5281/zenodo.3759876

11 minutes ago

First release of the PHENIX DAP site

Zenodo - GitHub panel - published release browser



April 21, 2020

PhenixCollaboration/web: First release of the PHENIX DAP site

Maxim Potekhin; Ron Belmont; amolhj

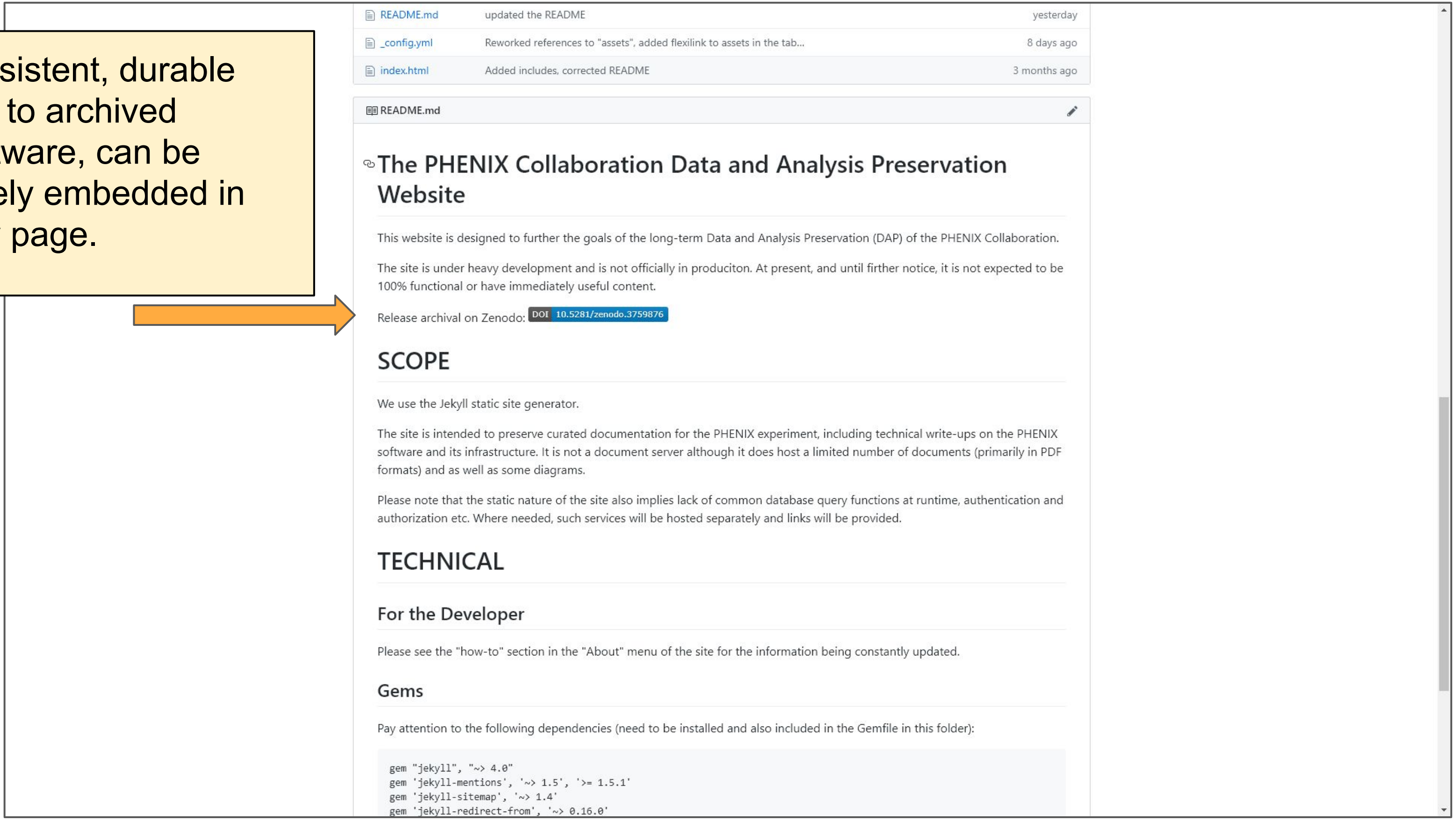
This is the first release of the PHENIX DAP website

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.



README.md	updated the README	yesterday
_config.yml	Reworked references to "assets", added flexilink to assets in the tab...	8 days ago
index.html	Added includes, corrected README	3 months ago

README.md

The PHENIX Collaboration Data and Analysis Preservation Website

This website is designed to further the goals of the long-term Data and Analysis Preservation (DAP) of the PHENIX Collaboration.

The site is under heavy development and is not officially in produciton. At present, and until firther notice, it is not expected to be 100% functional or have immediately useful content.

Release archival on Zenodo: DOI 10.5281/zenodo.3759876

SCOPE

We use the Jekyll static site generator.

The site is intended to preserve curated documentation for the PHENIX experiment, including technical write-ups on the PHENIX software and its infrastructure. It is not a document server although it does host a limited number of documents (primarily in PDF formats) and as well as some diagrams.

Please note that the static nature of the site also implies lack of common database query functions at runtime, authentication and authorization etc. Where needed, such services will be hosted separately and links will be provided.

TECHNICAL

For the Developer

Please see the "how-to" section in the "About" menu of the site for the information being constantly updated.

Gems

Pay attention to the following dependencies (need to be installed and also included in the Gemfile in this folder):

```
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)

Zenodo - final notes

- A drop-in replacement for DocDB and as such can fulfill immediate needs of the EIC community
- ...but with more functionality, future proof and interesting new features
  - Can store almost any type of data and this may be helpful for the EIC YR
  - Obviously not a production resource i.e. we won’t store datasets on Zenodo
  - However consider a versioned set of histograms as an example
  - Durable permalinks
  - API
- Search capabilities are significant and they will further improve
- An “EIC Software” community on Zenodo may be a useful thing to have (papers, conference presentations etc, good PR and a solid way to reference materials)
- A software group advice to the community?